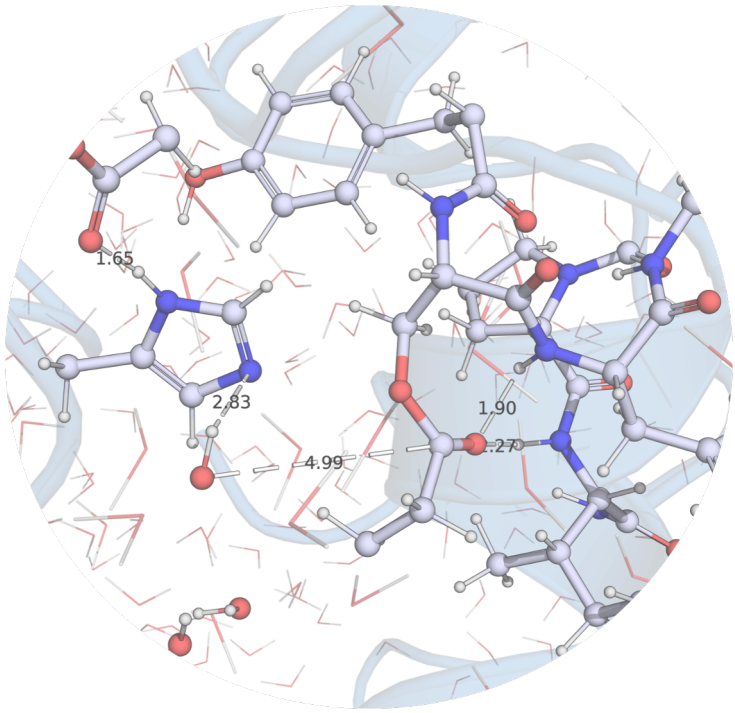


Understanding and Optimizing Extremozymes for Industrial Processes

Matilde F. Viegas
Programa Doutoral em Química
Departamento de Química e Bioquímica
2022

Orientador
Pedro Alexandrino Fernandes, Professor Associado, FCUP

Coorientador
Maria João Ramos, Professora Catedrática, FCUP

Acknowledgements

I would like to thank, first of all, my supervisor Professor Pedro Alexandrino Fernandes, for being so kind, generous, and supportive throughout this entire four-year-long journey. I would also like to thank my co-supervisor, Professor Maria João Ramos, for being an extraordinary role model and seeing the best qualities in each of us. Together, their joyous and positive outlook has inspired us to follow our curiosities and be our best selves. I feel lucky to have been able to carry out this project in such a warm work environment where failure is seen as an opportunity for growth and improvement. Isn't it what science is all about?

I would also like to thank all of my colleagues at the Theoretical and Computational Biochemistry group who taught me so much, particularly to Rui Neves, who always made the time to answer all of my questions with an enthusiastic spirit.

None of this would have been possible without the support of my mother, Mafalda, and of my closest friends, who soothed my doubts and provided encouragement when I most needed it.

Finally, I'm grateful for FCT for my PhD grant SFRH/BD/ 131968/2017.

Thank you all, it does take a village.

Abstract

This thesis lays out the current climate crisis we are facing and the crucial role of biofuels and green chemistry as tools to diminish greenhouse gas emissions and the toll of pollution. Enzymes have been at the forefront of green chemistry as the highest standard for environmentally-friendly catalysis. Yet, despite their immense power and potential, enzymes stand as undomesticated entities, as humans still struggle to make use of their abilities in their favor.

The use of computational chemistry in studying enzymatic catalysis has made strides in the atomistic understanding of enzymes and their catalytic mechanisms. In this thesis, we applied QM/MM methodologies, Homology Modeling and Protein-Protein Docking for the study of enzymes with potential for biomass processing and biofuel production, aiming to be part of the solution.

This thesis begins by exploring the realm of fatty acid synthesis: firstly, with the review “Animal Fatty Acid Synthase: A Chemical Nanofactory,” which provides the current state-of-art on type I fatty acid synthase (FAS) systems - explaining its catalytic mechanisms and its potential for therapeutics and industrial applications. Secondly, the first complete human Type I FAS model with its fourteen active domains is introduced. This structure was solved using homology modeling and a HADDOCK protein-protein docking server. The study of the type I FAS is further explored with the proposal of a QM/MM of the thioesterase domain, responsible for the cleavage and release of the synthesized fatty acids, for future catalytic mechanism studies using QM/MM ONIOM methods. Multidomain polyketide synthase (PKS) systems, similarly to Type I FAS, are megasynthases, and they are responsible for producing polyketides in bacteria, fungi, plants, and a few animal lineages. The review “Engineering of PKS Megaenzymes – A Promising Way to Biosynthesize High-Value Active Molecules” explores the potential of PKS enzymes.

The following chapter focuses on thermophilic enzymes with an initial review on thermophilic enzymes and their application in greener industries “Thermophilic Enzymes: The Key for Greener Industrial Processes.” We then introduce a glucuronoyl esterase (GE) enzyme with thermoresistant properties whose catalytic activity contributes to the degradation of lignocellulosic biomass – needed for biomass processing industries and biofuel production. GE’s catalytic mechanism was studied in detail using QM/MM ONIOM methods, resulting in the characterization of the ten stationary points of its reaction mechanism, alongside potential mutations that could be further explored to improve the enzyme’s catalytic rate.

The scope of the thesis extends to the study of snake venom: the secreted Phospholipase A2 enzyme found in snake venom with the published work “Structural, enzymatic and pharmacological profile of AplTX-II - A basic $sPLA_2$ (D49) isolated from the *Agkistrodon piscivorus leucostoma* snake venom”. For this study, we have resolved the structure of AplTX-II with the docked synthetic substrate NOBA (4-nitro-3-octanoyloxy benzoic acid) by homology modeling and performed MD simulations with explicit solvent.

Science communication was a key aspect during the four years of the Doctoral Program in Chemistry, emphasizing science education and outreach. Chapter 6 encompasses two articles published in the magazine Ciência Elementar, aimed at the Portuguese general public and high-school students, alongside a QM/MM protocol taught at the 2019 Dynapeutics International Summer School in San Sebastián, Basque Country (Spain), and a workbook devised for the Bioinformatic course taught at Faculty of Sciences of University of Porto and available on the Appendix section.

Keywords: Extremozymes, Fatty Acids, Fatty Acid Synthase, Glucuronoyl Esterase, Snake Venom, Science Communication, QM/MM, ONIOM, Homology Modeling, Molecular Dynamics

Resumo

Esta tese expõe a crise climática que enfrentamos, e o papel crucial dos biocombustíveis e da química verde como ferramentas para diminuir as emissões de gases de efeito estufa e o impacto da poluição. As enzimas estão na vanguarda da química verde como o mais alto padrão para catálise enzimática. Porém, apesar de seu imenso poder e potencial, as enzimas são ainda um grande mistério.

A compreensão atomística das enzimas e seus mecanismos catalíticos progrediu com os avanços e uso da Química Computacional no estudo da catálise enzimática. No contexto desta tese, forma usadas metodologias QM/MM, Modelação por Homologia e Ancoragem Proteína-Proteína no estudo de enzimas com potencial aplicabilidade no processamento de biomassa e na produção de biocombustíveis.

Esta tese começa por explorar o campo da síntese de ácidos gordos: primeiramente, com a *review* "*Animal Fatty Acid Synthase: A Chemical Nanofactory*", que estabelece o actual estado-da-arte em sistemas sintase de ácidos gordos (FAS) tipo I. São explicados os mecanismos catalíticos da FAS e a sua aplicabilidade terapêutica e industrial. Em segundo lugar, é apresentado o primeiro modelo completo de FAS tipo I humana e dos seus catorze domínios activos. A estrutura foi resolvida através de modelação por homologia em conjunto com ancoragem proteína-proteína utilizando o *software* HADDOCK. O estudo da FAS tipo I é ainda reforçado com a proposta de um modelo QM/MM do domínio Tioesterase, responsável pela clivagem e libertação dos ácidos gordos sintetizados. O presente modelo será útil para futuros estudos do mecanismo catalítico da Tioesterase com métodos QM/MM ONIOM. À semelhança dos sistemas FAS tipo I, os sistemas sintase de policetídeos (PKS) são mega-sintases com vários domínios, e são responsáveis pela produção de policetídeos em bactérias, fungos, plantas e alguns animais. A *review* "*Engineering of PKS Megaenzymes - A Promising Way to Biosynthesize High-Value Active Molecules*" explora o potencial das enzimas PKS.

O seguinte capítulo tem como foco as enzimas termofílicas, iniciando com uma *mini review* sobre enzimas termofílicas e a sua aplicação em indústrias verdes: "*Thermophilic Enzymes: The Key for Greener Industrial Processes*." Em seguida, é apresentado o estudo QM/MM da enzima glucuronoil esterase (GE) cuja actividade catalítica e propriedades termorresistentes contribuem para a degradação da biomassa lignocelulósica – útil para indústrias de processamento de biomassa e produção de biocombustíveis. O mecanismo catalítico da GE foi estudado usando métodos QM/MM ONIOM, tendo sido feita a caracterização dos dez pontos estacionários de seu mecanismo de reacção, juntamente com potenciais mutações que poderão ser estudadas com o objectivo de melhorar a taxa catalítica da enzima.

A tese estende-se ainda ao estudo da enzima fosfolipase A2 secretada, encontrada em veneno de cobra, com o trabalho publicado "*Structural, enzymatic and pharmacological profile of AplTX-II - A basic sPLA$_2$ (D49) isolated from the* Agkistrodon piscivorus leucostoma *snake venom*". Neste estudo foi resolvida a estrutura de AplTX-II com o substrato sintético NOBA (ácido 4-nitro-3-octanoiloxibenzóico), usando modelação por homologia e simulações de MD realizadas em solvente explícito.

A comunicação científica foi um aspecto fundamental durante os quatro anos do Programa Doutoral em Química, com ênfase na educação e divulgação científica. O capítulo 6 engloba dois artigos publicados na revista Ciência Elementar, dirigida a estudantes do ensino secundário e público português, a par com um protocolo QM/MM que foi leccionado na Escola Internacional de Verão Dynapeutics 2019 em San Sebastián, País Basco (Espanha) e da Sebenta para a cadeira de Bioinformática leccionada na Faculdade de Ciências da Universidade do Porto, disponível na secção Anexo.

Palavras-chave: Extremozimas, Ácidos Gordos, Sintase de Ácidos Gordos, Glucuronoil Esterase, Veneno de Cobra, Comunicação Científica, QM/MM, ONIOM, Modelação por Homologia, Dinâmica Molecular

Index

ACKNOWLEDGEMENTS ..... **I**

ABSTRACT ..... **II**

RESUMO ..... **IV**

INDEX ..... **VI**

INDEX OF FIGURES ..... **X**

INDEX OF SCHEMES ..... **XIV**

INDEX OF TABLES ..... **XV**

LIST OF ABBREVIATIONS ..... **XVI**

LIST OF PUBLICATIONS ..... **XIX**

CHAPTER 1. ENZYMES AS THE KEY TO A SUSTAINABLE FUTURE **1**

1.1. Climate change, a century-long issue 2

1.2. Onwards with Biofuels 3

1.2.1. Biodiesel *3*

1.2.2. Bioethanol *5*

1.3. Enzymes as a potential solution for greener industries 6

1.3.1. Fatty Acid Synthase as a biofuel production asset *8*

1.3.2. Mesophilic vs extremophilic enzymes in industrial processes *11*

1.3.3. Extremozymes structural properties that make up their armor *12*

1.3.4. Enzymes for biomass processing *14*

1.3.5. Thermophilic Glucuronoyl Esterase *15*

1.4. Enzymatic engineering and yield optimization methods 16

1.5. Enzymes as therapeutics: the case of snake venom 19

1.6. Science communication and outreach 20

1.7. Outline of the thesis 21

CHAPTER 2. THEORY BEHIND COMPUTATIONAL METHODS FOR THE STUDY OF ENZYME CATALYSIS **24**

2.1. Homology Modelling 25

2.2. Protein-Protein Docking 27

2.3. Molecular Mechanics 29

2.3.1. Force Field *30*

2.3.2. Molecular Dynamics ... *35*
2.4. Quantum Mechanics ... 39
2.4.1. Density Functional Theory (DFT) ... *42*
2.4.2. Exchange-correlation density functionals ... *43*
2.4.3. Basis sets ... *44*
2.5. Hybrid QM/MM Methods ... 46
2.5.1. Subtractive ONIOM scheme ... *47*
2.5.2. Mechanical embedding, electrostatic embedding and link atoms ... *48*
2.5.3. QM/MM geometry optimization ... *50*
2.5.4. Best practices for the use of hybrid QM/MM methodology ... *51*
2.5.5. How to study the catalytic mechanism of an enzyme using QM/MM? ... *52*
Preparation of the system ... 52
Establishing QM and MM layers ... 54
Choosing the level of theory ... 55
How to explore the reaction coordinate? ... 56
Establishing the final reaction profile ... 57
CHAPTER 3. FATTY ACID SYNTHASE ... 58
3.1. Fatty Acid Synthase: an introduction ... 58
3.2. Modeling of Human Fatty Acid Synthase and *in Silico* Docking of Acyl Carrier Protein Domain and Its Partner Catalytic Domains ... 61
3.2.1. Abstract ... 63
3.3. Modeling of Type I FAS Thioesterase Domain and preparation of QM/MM model for upcoming catalytic mechanism studies ... 73
3.3.1. Abstract ... 74
3.3.2. Introduction ... 75
TE domain's role in animal FAS ... *75*
Structure Overview ... *76*
Substrate Binding Site and Substrate Specificity ... *77*
The Catalytic Mechanism ... *79*
3.3.3. Methods ... 81
Acyl-Enzyme Model Preparation ... 81
Molecular Dynamics ... 84
3.3.4. Results ... 84
3.3.5. Future steps ... 86
3.4. Outcomes and Future Prospects ... 88
3.4.1. Animal Fatty Acid Synthase: a Chemical Nanofactory ... 89

3.4.1.1. Abstract *90*
3.4.2. Engineering of PKS megaenzymes: a promising way to biosynthesize high-value active molecules 92
3.4.2.1. Abstract *93*
CHAPTER 4. STUDY OF GLUCURONOYL ESTERASE 113
4.1. Thermophilic Enzymes: The Key for Greener Industrial Processes 115
4.1.1. Abstract 116
4.2. QM/MM Study of the Reaction Mechanism of the Thermophilic Glucuronoyl Esterase for Biomass Treatment 128
4.2.1. Abstract 129
4.2.2. Introduction 130
4.2.2.1. The biocatalytic potential of extremophiles for greener processes *130*
4.2.2.2. Glucuronoyl Esterase: an LCC degrading machine *130*
4.2.2.3. Nature's own catalysts as industrial workhorses *134*
4.2.3. Computational Methods 136
4.2.3.1. Model Preparation *136*
4.2.3.2. QM/MM Model *139*
4.2.3.3. QM/MM calculations and reaction coordinates *140*
4.2.4. Results and discussion 141
4.2.4.1. Water: the key for the catalysis *141*
4.2.4.2. First Reaction Stage: Formation of the acyl-serine enzyme intermediate *143*
4.2.4.3. Second Reaction Stage: Deacylation and Serine Regeneration *146*
4.2.4.4. Energetic profile for the GE catalytic mechanism *147*
4.2.4.5. Transition-State Macromolecular Dipole Stabilization *149*
4.2.4.6. Comparison with members of the CE15 family - active site rigidity and substrate binding *151*
4.2.5. Conclusions 155
4.3. Outcomes and Future Prospects 158
CHAPTER 5. SNAKE VENOM AS A RESERVOIR FOR THERAPEUTIC AGENTS 160
5.1. Structural, Enzymatic and Pharmacological profile of AplTX-II - A Basic sPLA$_2$ (D49) isolated from the *Agkistrodon piscivorus leucostoma* snake venom 161
5.2.1. Abstract 162
5.3. Outcomes and Future Prospects 177
CHAPTER 6. SCIENCE COMMUNICATION 178
6.1. Biocombustíveis, serão a solução? 178
6.2. Proteínas Virais no *Protein Data Bank* 185

6.3. Protocol for QM/MM Workshop at 2019 Dynapeutics Summer School 190
6.6. Outcomes and Future Prospects 195

CHAPTER 7. FINAL REMARKS 196

CHAPTER 8. APPENDIX 198

8.1. Evaluation of a flavonoids library for inhibition of pancreatic a-amylase towards a structure–activity relationship 199
8.2. Animal Fatty Acid Synthase: a Chemical Nanofactory 212
8.3. Sebenta de Bioinformática Molecular 265

REFERENCES 318

Index of Figures

Figure 1| Representation of the Type I FAS. A) Representation of the animal FAS homodimer head-to-head model, as obtained from homology modeling [55]. B) a representation of the dodecameric assembly of the barrel-shaped fungal FAS 10
Figure 2| Lignocellulosic biomass. Illustrative and schematic representation of lignocellulosic biomass structure composed by cellulose, hemicellulose and lignin, where the covalent linkages of lignin-carbohydrate complexes (LCC) are found. Biocatalysts for the processing of LCC contribute to the reduction of costs and pollution. Illustration © Mafalda Salgueiro. 15
Figure 3| Schematic representation of the four bonded energy terms. r represents the bond length between atoms A and B; θ is the bond angle between atoms A, B, and C; φ is the dihedral torsion between atoms A, B, C and D, and the φ is the angle formed by two non-concurrent vectors in the plan. Scheme © Mafalda Salgueiro. .. 31
Figure 4| Balance is key for a successful MD simulation. MD simulations depend on a fine balance between the length of the simulation, the size of the system and the accuracy of the method. Illustration © Mafalda Salgueiro. 37
Figure 5| Link atoms. Link atoms are placed between the atom of the low layer and the atom of the high layer, and covalently bound to the last one (high layer). The goal of the link atom is to cap the free valence of the active part in the high layer. Illustration © Mafalda Salgueiro. 50
Figure 6| The TE domain of the animal (human) FAS [55]. A) Three-dimensional surface view of the TE domain within the human FAS monomer. B) TE secondary structure elements and active site region (within the red outlined box). The TE-substrate complex is taken from a computational model based on the crystallographic structure of the Homo sapiens TE domain (PDB ID: 2PX6) with a modelled palmitic acid molecule. C) Close-up of the active site cavity. The catalytic residues and the product are depicted in ball-and-stick representation. 76
Figure 7| Structure of the binding site of the TE domain of the animal (human) FAS [55] A) Close-up three-dimensional surface view of the TE domain within the human FAS monomer (PDB ID: 2PX6). The active site cavity shows two distinct binding sites:

Site 1, for short-length C chains, located at the interface between the two TE subdomains; and Site 2 for the C16-C18 substrates. The Helix Flap (PDB ID: 3TJM) highlighted in grey cartoon is responsible for the formation of the deep groove-tunnel Site 2 that harbors the C16-C18 substrate. The ligand Orlistat is represented in yellow ball-and-stick representation, with its peptidyl moiety protruding into the Site 1 resembling the behavior of short-length carbon chains. The C18 polyunsaturated fatty acid called methyl γ- linolenylfluorophosphonate (MGLFP, PDB ID: 3TJM) ligand is represented in grey ball-and-stick representation. For clarity, the active site residues are absent, except for the indication of the position of the Ser2308 that covalently binds the substrates. B) Cartoon representation of both the PDB ID: 2PX6 and 3TJM, aligned, displaying the helix flap in its disordered form (blue, PDB ID: 2PX6) and ordered form (grey, PDB ID: 3TJM). 77

Figure 8| Representation of the TE domain's active site, with MGLFP and Orlistat ligands bound to Ser2308...... 82

Figure 9| QM/MM model of the TE-Palmitoyl intermediate. Transparent cartoon representation of the TE protein, as well as the transparent stick representation of water molecules, depict the considered MM layer. The atoms represented in ball and stick belong to the QM layer. The QM layer includes the catalytic triad Ser2308, His2481, Asp2338 and the palmitoyl moiety, alongside Tyr2309, Ile2250 that stabilize the oxyanion hole. Tyr2307, Pro2249 and Gly2310 establish a network of hydrogen bonds that contribute to the stabilization and rigidity of the active site and therefore should be included in the QM layer. 85

Figure 10| Representation of the *Tt*GE enzyme, obtained from the structure PDB ID: 4G4J. Catalytic residues are shown in stick representation (colored by element, yellow C atoms), and the docked substrate MGPLGN is shown in stick representation, colored by element with purple C atoms. The C6 carbon atom of the substrate MGPLGN is represented with a purple sphere. 132

Figure 11| Overall QM/MM models for Stage 1 and Stage 2 of the reaction mechanism. Transparent cartoon representation of the GE protein, as well as the transparent stick representation of water molecules, depict the considered MM layer. The atoms represented in ball and stick belong to the QM layer. 140

Figure 12| Water molecules that can be found in the immediacy of the active site (within 5 Å of both carbon C6 of the glycosyl-enzyme and nitrogen Nε of His346), every 1 ns over the course of an unrestrained 40 ns cMD simulation. Water molecules are represented in red spheres, surrounding the active site residues, in yellow colored stick representation, and the glycosyl MGP, in purple colored stick representation. Carbon C6 of the glycosyl-enzyme and nitrogen Nε of the His346 are highlighted in sphere representation. 142
Figure 13| Stationary points of the *Tt*GE reaction. QM layer represented in ball and stick, and the substrate colored in purple. In the INT2' stage, the nucleophilic bulk water molecule is highlighted with a red-colored circle. In dashes, the most relevant Interatomic distances that were followed along the QM/MM calculations, obtained at the B3LYP/6-31G(d):FF99SB level. 145
Figure 14| Gibbs Energy profile of the global GE-catalyzed carbohydrate hydrolysis, at the DLPNO-CCDS(T)/CBS:FF99SB//B3LYP/631G(d):FF99SB level of theory. The interruption between the INT2 and INT2' states, of both Stages 1 and 2, was generated by the removal of the free LGN leaving group. 148
Figure 15| Per-residue activation energy contribution as a function of the difference in the distances of the centre of mass of the sidechain (or the charged group for the charged residues) of the residue to the ester carbonyl oxygen of the substrate MGPLGN and Nε of His346, in Å. Larger circles represent bulkier amino acids. Positively charged, negatively charged, polar, and hydrophobic residues are colored, respectively, in blue, red, green, and gray. The energy contributions were evaluated at the ONIOM(B3LYP/6-311+G(2d,2p):AMBER) level. B) Representation of the residues within 10 Å of the active site of TtGE. Positively charged, negatively charged, polar, and hydrophobic residues are colored, respectively, in blue, red, green, and grey. Labelled residues lead to the most significant changes in the activation energy. The dashed grey line separates residues closer to ester carbonyl oxygen of MGPLGN than to NεH of His346. 151
Figure 16| 2D Diagrams of the H-Bond interactions between the ligand and neighboring residues present at the Active Site, displaying the predominance of interactions established between the uronic acid moiety and the protein. A) the Crystal Structure of TtGE (PDB ID:4G4J) and its co-crystalized MCU ligand; B) At the

Active Site of the TtGE with the MGPLGN ligand used for the study of the Catalytic Reaction at Stage 1.....154
Figure 17| Electrostatic Surface Potential representation of fungal and bacterial GEs. A. Fungal TtGE (PDB ID: 4G4J) with docked substrate; B. Fungal CuGE with docked substrate (PDB ID:6RV9, chain B). C) Bacterial OtGE with docked substrate (PDB ID: 6T0I).....154

Index of Schemes

Scheme 1| Catalytic mechanism proposed for FAS TE domain. Proposed mechanistic pathways for the TE domain catalytic mechanism. ("R" stands for the acyl moiety of C16 that will be cleaved by the TE domain. The residues are numbered according to the human TE domain). The Gibbs activation energy of 18.1 kcal·mol^{-1} was derived from the k_{cat} value of 0.32 s^{-1} (25 °C) [239] using the Transition State Theory. 80

Scheme 2| TE domain products: palmitic acid (C16) and stearic acid (C18). ... 83

Scheme 3| MGPLGN substrate representative of LCC, representative of the most recalcitrant fraction of LCC – α ester – designed based on the MCU substrate of the *Tt*GE PDB ID: 4G4J X-ray structure [111]. MGPLGN is formed by a hemicellulose carbohydrate 4-O-methyl-β-D-glucopyranuronate (MGP) and a lignin (LGN) moiety (highlighted in red shade). The ester group is represented in red lines. This substrate is used as a representative minimal LCC in the study of the catalysis of *Tt*GE. 131

Scheme 4| Schematic representation of the proposed canonical mechanism for esterase enzymes [249]. The initial substrate and its progression in the reaction mechanism is highlighted in blue. 133

Scheme 5| Schematic representation of Glucuronoyl Esterase reaction mechanism. The initial substrate and its progression in the reaction mechanism is highlighted in blue. 149

Index of Tables

Table 1| Experimental proposal for the study of the behavior of the helix flap and its interaction with different-sized substrates, the binding of ACP to the TE domain of FAS, throughout the many stages of the elongation cycle. 87

List of Abbreviations

3D	Three-Dimensional
ACP	Acyl Carrier Protein
AHAS	Acetohydroxy Acid Synthase
AIR	Ambiguous Interaction Restraints
AMBER	Assisted Model Building with Energy Refinement
BLAST	Basic Local Alignment Search Tool
CAZy	Carbohydrate-Active Enzymes
CE	Carbohydrate Esterase
CGTO	Contracted Gaussian Type Orbitals
CMN	*Corynebacterium*, *Mycobacterium*, and *Nocardia* group bacteria
***Cu*GE**	*Cerrena unicolor* glucuronoyl esterase
DFT	Density Functional Theory
DH	Dehydratase domain
E.C.	Enzyme Commission number
EE	Electrostatic Embedding
ER	Enoyl Reductase domain
ESP	Electrostatic Potential
FAS	Fatty Acid Synthase
FAEE	Fatty Acid Ethyl Esters
FAME	Fatty Acid Methyl Esters
FDA	USA Food and Drug Administration
GAFF	Generalised AMBER Force Field
GE	Glucuronoyl Esterase
GGA	Generalized Gradient Approximation
GOLD	Genetic Optimization Ligand Docking
GTO	Gaussian Type Orbitals
HF	Hartree-Fock
hFAS	human Fatty Acid Synthase
INT	Reaction Intermediate

IRC	Intrinsic Reaction Coordinate
K_{cat}	First order rate constant
K_m	Michaelis-Menten constant
KR	β-Ketoacyl Reductase domain
KS	β-Ketoacyl Synthase domain
LCC	Lignin-Carbohydrate-Complexes
LDA	Local Density Approximation
LGN	Lignin
MAT	Malonyl Acetyl Transferase domain
MCU	4-*O*-methyl-β-*D*-glucopyranuronate
MD	Molecular Dynamics
ME	Mechanical Embedding
MGLFP	Methyl γ-linolenylfluorophosphonate
MM	Molecular Mechanics
NCBI	National Center for Biotechnology Information
NMR	Nuclear Magnetic Resonance
NOBA	4-nitro-3-octanoyloxy benzoic acid
NPT	Isobaric Isothermal Ensemble
NVT	Canonical Ensemble
ONIOM	Our own N-layered Integrated molecular Orbital and molecular Mechanics
***Ot*GE**	*Opitutus terrae* glucuronoyl esterase
PBC	Periodic Boundary Conditions
PDB	Protein Data Bank
PES	Potential Energy Surface
PKS	Polyketide Synthase
PME	Particle Mesh Ewald
PNS	4'-phosphopantetheine
QM	Quantum Mechanics
QM/MM	Quantum Mechanics / Molecular Mechanics
RESP	Restrained Electrostatic Potential
RMSd	Root Mean Square deviation

SPL	Serine-Palmitoyl modified residue
sPLA$_2$	secreted Phospholipase A_2
STO	Slater Type Orbitals
TE	Thioesterase
TIP3P	Transferable Intermolecular Potential 3P
TS	Transition State
***Tt*GE**	*Thermothelomyces thermophila* glucuronoyl esterase

List of Publications

This thesis results from the following publications, and two manuscripts currently under submission:

Viegas, M. F., Neves, R. P. P., Ramos, M. J., & Fernandes, P. A. (2021). ***QM/MM Study of the Reaction Mechanism of the Thermophilic Glucuronoyl Esterase for Biomass Treatment.*** (Manuscript under revision by Catalysis Science & Technology).

Oliveira, A. L., **Viegas M. F.**, Da Silva S. L., Soares, A. M., Ramos M. J. & Fernandes, P. A. (2021). ***The Chemistry of Snake Venom and its Medicinal Potential.*** (Manuscript under revision by Nature Reviews Chemistry).

Neves, R. P. P., Ferreira, P., Medina, F. E., Paiva, P., Sousa, J. P. M., **Viegas, M. F.,** Fernandes, P. A., & Ramos, M. J. (2021). ***Engineering of PKS Megaenzymes—A Promising Way to Biosynthesize High-Value Active Molecules.*** In Topics in Catalysis. Springer Science and Business Media LLC. doi:10.1007/s11244-021-01490-5

Paiva, P., Medina, F. E., **Viegas, M**., Ferreira, P., Neves, R. P. P., Sousa, J. P. M., Ramos, M. J., & Fernandes, P. A. (2021). ***Animal Fatty Acid Synthase: A Chemical Nanofactory.*** In Chemical Reviews (Vol. 121, Issue 15, pp. 9502–9553). American Chemical Society (ACS). doi:10.1021/acs.chemrev.1c00147

Viegas, M., Ramos, M. J., & Alexandrino Fernandes, P. (2021). ***Thermophilic Enzymes: The Key for Greener Industrial Processes.*** In U.Porto Journal of Engineering (Vol. 7, Issue 3, pp. 13–23). University of Porto. doi:10.24840/2183-6493_007.003_0002

Resende, L. M., Almeida, J. R., Guaraca-Medina, T. A., **Viegas, M. F.**, Soares, A. M., Ramos, M. J., Fernandes, P. A., Marangoni, S., & Da Silva, S. L. (2021). ***Structural,***

enzymatic and pharmacological profiles of ApiTX-II - A basic sPLA$_2$ (D49) isolated from the Agkistrodon piscivorus leucostoma snake venom. In International Journal of Biological Macromolecules (Vol. 175, pp. 572–585). Elsevier BV. doi:10.1016/j.ijbiomac.2021.01.187

Pinto, A. V. P., Oliveira, A. N. L. O., Jerves, C. J., Mendonça, F. L. de M., Coutinho, I. F. M. C., Coimbra, J. T. S. C., Sousa, J. P. M. S., Cabrera-Trujillo, J. J. C.-T., Teixeira, L. T., **Viegas, M. F.**, Ferreira, P. F., Paiva, P. P., Fonseca, R. F., Neves, R. P. P. N., Magalhães, A. L. M., Fernandes, P., Ramos, M. J. (2020). ***Proteínas Virais no Protein Data Bank***. In Revista de Ciência Elementar (Vol. 8, Issue 2). ICETA. doi:10.24927/rce2020.022

Proença, C., Freitas, M., Ribeiro, D., Tomé, S. M., Oliveira, E. F. T., **Viegas, M. F.**, Araújo, A. N., Ramos, M. J., Silva, A. M. S., Fernandes, P. A., & Fernandes, E. (2019). ***Evaluation of a flavonoids library for inhibition of pancreatic α-amylase towards a structure–activity relationship.*** In Journal of Enzyme Inhibition and Medicinal Chemistry (Vol. 34, Issue 1, pp. 577–588). Informa UK Limited. doi:10.1080/14756366.2018.1558221

Viegas, M., Ramos, M. J., & Fernandes, P. (2018). ***Biocombustíveis, serão a solução?*** In Revista de Ciência Elementar (Vol. 6, Issue 2). ICETA. doi:10.24927/rce2018.043

Viegas, M. F., Neves, R. P. P., Ramos, M. J., & Fernandes, P. A. (2018). ***Modeling of Human Fatty Acid Synthase and in Silico Docking of Acyl Carrier Protein Domain and Its Partner Catalytic Domains.*** In The Journal of Physical Chemistry B (Vol. 122, Issue 1, pp. 77–85). American Chemical Society (ACS). doi:10.1021/acs.jpcb.7b09645

CHAPTER 1. Enzymes as the key to a sustainable future

Older generations were able to enjoy and do whatever they wanted without worrying about climate refugees and extreme climate disasters. But now the world around us is deteriorating beyond a point of natural recovery, and we are not going to be able to inhabit a world where nature is healthy and in balance. It's as unfair to ecosystems as it is to humans.

Our generation holds a kind of consciousness that is not based on monetary gain or on new ways of profiting from lands, forests, rivers, seas, and people. We are pushing for a complete breakthrough of sensible and wise solutions that ensure the continuation of ecosystems and peaceful societies. For this, human civilization needs to make drastic changes in its value system; it needs to mature.

Johnson, A. E., Wilkinson, K. K., & Bastida, X. (2021). Calling In. In All we can save: Truth, courage, and solutions for the Climate Crisis. essay, One World.

1.1. Climate change, a century-long issue

Until the 19th century, societies relied on renewable energy for heat and mechanical power provided by the annual cycle of plant photosynthesis. This limited amount of energy curbed population and industry growth over centuries. In the early 19th century, the First Industrial Revolution brought readily available energy through the usage of coal in steam engines, independent of plant growth and photosynthesis [12]. The industrialization that followed restructured western societies, luring whole families from rural areas into cities and towns, which caused more frequent epidemics due to lack of hygiene practices and installed poverty [3].

The Second Industrial Revolution began in the late 19th century in Britain, the United States, and other now-developed countries. It was triggered by inventions such as the light bulb by Thomas Edison, which led to a phenomenon known as electrification. Over the 20th century, like the railroad and telephone systems, electricity became accessible through the connection of electrified niches, feeding a growing grid that brought electric power to entire nations [4]. As electric power became available to an increasing number of people, its consumption increased, causing a simultaneous shift of energy sources: plant photosynthesis was still the prime energy source, but its preferred form was now coal, where energy was harnessed and accumulated over geological age [2].

Since the Industrial Revolution until now, the propelling force for the energy industry is a society based on consumption and demand, but current circumstances, such as *i)* population growth, *ii)* unprecedented need for energy as economies in Asia, Latin America, and Africa emerged, and *iii)* Global Warming, alongside the *iv)* shortage of fossil fuels, forced action towards a more sustainable approach to energy production [5]. As a result, world leaders introduced specific and tangible goals, firstly, with the Kyoto Protocol in 1997 [6,7], and later with its follow-up Paris Agreement [8] signed in 2015, which meant to keep global temperatures "well below" 2.0 ºC above pre-industrial times and "endeavor to limit" them even more, to 1.5 ºC above pre-industrial times. However, as these goals weren't met in the consequent years, COP26

concluded in Glasgow with nearly 200 countries agreeing on the Glasgow Climate Pact to keep "1.5 °C" alive and finalize the outstanding elements of the Paris Agreement.

Despite the ambitious guidelines established, the lack of commitment by international governments and the private sector has resulted in movements advocating for ecological responsibility at an individual level, encouraging recycling, reducing meat consumption, and reducing plastic use. This concerted effort has also impacted energy industries, as we are witnessing the turn towards an energy system based on sustainable, renewable practices. These practices center on biofuels (oil-derived biodiesel, bioethanol produced by fermentation processes, and biogas) made from combustible sources, employing enzymatic catalysts.

1.2. Onwards with Biofuels

In recent decades, a growing interest has been drawn to biofuels which are originated through the processing of renewable feedstocks such as cellulosic materials, oil-rich plant seeds, waste plant biomass, food stocks, to name a few. These are considered renewable stocks because plants require carbon dioxide for growth and they release it back to the atmosphere during combustion, resulting in net-zero carbon dioxide emissions. This renewable energy cycle is beneficial to the environment, as it does not emit additional greenhouse gases to the atmosphere. However, there are many issues yet to be tackled regarding biofuel: *i)* its production, yet in its developing and optimizing stages, and *ii)* its compatibility with standard car engines, built for traditional fuels like diesel and gasoline, and *iii)* distribution infrastructure [9].
Considering the scope of the present thesis, a brief introduction to biodiesel and bioethanol biofuels will follow:

1.2.1. Biodiesel

At the 1900 World Exhibition in Paris, Dr. Diesel exhibited an engine running 100% peanut oil [10], laying the ground stone for society fueled with clean, renewable, locally

grown fuel [10]. Among the different biofuels, biodiesel has been the most successfully produced. Biodiesel is methyl or ethyl ester of fatty acid obtained from virgin or waste vegetable oils (both edible and non-edible) and animal fat, and it is readily available at gas stations through biodiesel blends. A brief overview of the advantages of fatty-acid-based fuels includes high energy density, low hygroscopicity, miscibility with diesel fuels, compatibility with existing infrastructure, and reduced purification costs [11]. In addition, biodiesel exhaust emissions are less harmful to the environment when compared to the ones originating from fossil fuels, with fewer greenhouse gas emissions [9]. Production-wise, biodiesel fuels are appealing because they can be produced from sustainable stocks ranging from edible to non-edible oils [12], to waste animal fats, waste cooking oil [13], and from algae (microalgae and macroalgae) [14]. The wide range of prime-matter rounds up different sources, circumventing the possible scenario of resource depletion.

Industrial production of biodiesel is predominantly achieved through transesterification [15]. In this reaction, triglycerides from animal fats or vegetable oils react with an alcohol in the presence of a strong acid or base, resulting in the breakdown of ester bonds between fatty acids and the -OH groups of glycerol. The resulting free fatty acids make new ester bonds with alcohol molecules present in the transesterification mixture. The preferred method for industrial-scale transesterification is via alkaline catalysis. However, alkaline catalysis often results in a side reaction called saponification of free fatty acids, leading to a reduced volume of catalysts and products [16].

Depending on the type of alcohol, the resulting fatty acids will vary. If employing methanol, the preferred one for its cost, the resulting mixture will be Fatty Acid Methyl Esters, so-called FAME, and Glycerol. If ethanol is used instead of methanol, the mixture will be Fatty Acid Ethyl Esters (FAEE) and glycerol.

In the context of biodiesel production, particularly FAEE, the Fatty Acid Synthase (FAS) enzyme has sparked interest. Fatty acids produced by Type I (large multifunctional enzymes) or Type II (distinct, individualized enzymes) FAS enzymes can be converted into FAEEs through esterification with ethanol [17,18]. More on the

FAS system in subchapter 1.3.1 followed by chapter 3 of the present thesis, which delves into the human Type I FAS, firstly through the development of a homology model of the human FAS, and secondly, a preliminary approach to modeling of the thioesterase domain of FAS for future catalytic mechanism computational studies and enzymatic engineering.

1.2.2. Bioethanol

Bioethanol is a clean fuel produced from biomass or waste through the fermentation of carbohydrates units into alcohol.

$$C_6H_{12}O_6 \rightarrow 2\ C_2H_6O + 2\ CO_2$$

EQUATION 1 - GLUCOSE CONVERSION INTO ETHANOL AND CARBON DIOXIDE

Bioethanol feedstocks include starchy material, such as sugar cane and corn, and lignocellulosic biomass, mostly from agricultural and forest waste, and nonedible parts of food crops. Lignocellulosic matter is a more environmental-friendly option due to its availability, low cost, and non-competition with the food market as opposed to sugar cane and corn.

Depending on the feedstock used for bioethanol production, the type and extent of pretreatments will vary. For starchy materials, physical size reduction is sufficient to ensure suitable quality feedstock and easy access of enzymes that will hydrolyze carbohydrates into fermentable sugars [19]. Lignocellulose is energetically valuable as it is mainly composed of sugars from plant cell walls (75% of polysaccharides) [20]. However, the structural complexity of lignocellulosic biomass makes it harder to process for biofuel production than starch-based feedstocks: fermentable sugars are trapped inside of intertwined chains that make up lignocellulose and whose main structural units are cellulose, hemicellulose, and lignin. Cellulose is a crystalline structure of thousands of β(1-4)D-Glucose units [21], hemicellulose is a randomly distributed net of xylose and arabinose polymers, and lignin is a biopolymer made of an aromatic compound. Many types of physical, chemical, and biological pretreatments for lignocellulosic material are often combined to achieve maximum efficiency. Biological pretreatments employ microorganisms, particularly *fungi*, to

degrade the lignocellulosic components of the feedstock. These fungi types are brown rot, soft rot, and, the most preferred one, white-rot fungi [22–24], all of which have cellulolytic properties. The use of living organisms as catalysts requires no added chemicals, therefore proving itself as a sustainable and harmless alternative to chemical methods.
Moreover, chapter 4 explores the catalytic mechanism of Glucuronoyl esterase, a lignocellulose degrading enzyme found in thermophilic the fungus *Thermothelomyces thermophila*.

1.3. Enzymes as a potential solution for greener industries

Enzymes are key players in life's rhythm. They are responsible for catalyzing (accelerating) chemical reactions essential to life maintenance by lowering their activation energy, and they can perform this task without themselves being consumed by the reaction [25]. A textbook example is that of carbonic anhydrase, the enzyme responsible for the reversible hydration of carbon dioxide, allowing for a complete and fast transfer of CO_2 from the tissue, into the blood, to the alveolus [26]. In the case of fish, the uncatalyzed conversion of HCO_3^- to CO_2 may require more than 1 min to approach completion, yet the blood remains in the gills for only 1-3 s. It is the red blood cell carbonic anhydrase role to increase the rate of CO_2 hydration/dehydration reactions up to 25000 fold to allow conversions between HCO_3^- and CO_2 to reach near completion in milliseconds, enabling effective CO_2 excretion [27].

Humans were inconspicuous about the existence of enzymes until the nineteenth century when diastase (a mixture of amylases) was the first enzyme to be discovered 1833 [28], and Wilhelm Kühne coined the term enzyme in 1877 [29]. Nevertheless, the use of enzymes by humans has been around for millennia. The most ancient proof dates back to the seventh millennium B.C.; chemical analysis of ancient organics absorbed into pottery jars revealed a mixed fermented beverage of rice, honey, and fruit [30]. Five millennia ago, humans naively used rennet, a set of enzymes found in animal stomachs, to process milk into cheese [31]. A well-known case of the use of enzymes in transformation processes is that of *Saccharomyces cerevisiae*: it has been extensively

studied, used, and genetic and metabolically engineered for fermentation despite the existence of thousands of other microorganisms catalyzing glucose fermentation [32].

Besides *S. cerevisiae*'s relevance in food industries, we can find widespread use of enzymes in other industries markets whose prime-matter are biomolecules that can be degraded, modified, or produced by enzymes. These range from textile [33] to animal feed [34–36], paper [37–39], biofuel production [19,40,41], pharmaceutical [42], and others.

The diversity of enzymes existing in Nature is vast. Nearly 100 thousand enzyme structures have been resolved at the time of publication of this thesis and made available on Protein Data Bank (PDB) [43]. Many systems have been devised to classify enzymes; the most prevailing one is an international agreement, developed by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology [44], that classifies enzymes according to the type of reactions they catalyze. This classification divides enzymes into seven distinct classes and following subclasses, using an Enzyme Commission number (E.C.) with four digits. Enzymes, Nature's catalysts for a myriad of chemical reactions essential to life, offer high specificity and reaction rate enhancement - the ideal scaffold for rate-enhancing enzymatic engineering. The proven usefulness of enzymatic catalysis in industrial processes, combined with environmental and climate concerns of the past decades, has further fueled scientific and technological interest in these macromolecules with the drive to maximize their potential as viable greener alternatives for industrial processes.

Successful enzymatic catalysis in industrial contexts offers the possibility of reduced energy input, lower costs, less polluting processes, and allows for efficient product recovery at the end. The applicability of enzymatic catalyzers in industrial settings first began in the 1960s, however, its production is often met with hurdles. The global market for industrial enzymes was already valued at USD 5.5 billion and is expected to reach USD 7.0 billion by 2023 [45]. As cutting-edge technologies continue to be developed, improving the production and purification of enzymes in industrial

quantities, the global enzyme market is expected to expand even more over the following years.

1.3.1. Fatty Acid Synthase as a biofuel production asset

Enzymatic catalysts are a promising alternative to the industrial processes that prevail in the biofuel production industry. They bring several economic and environmental advantages, such as high selectivity, milder reaction temperatures, and no side reactions or polluting by-products. Amongst the most frequently used enzymes for biodiesel production are organisms belonging to the taxonomic orders *Candida*, *Pseudomonas*, and *Rhizopus* [46]. As the search for means to engineer and optimize fatty-acid output for biofuel production continues, the FAS system poses as a strong candidate. Fatty acids are central components for energy storage and biological membranes, therefore essential to life. This requirement has resulted in fatty acid biosynthesis systems being found in all kinds of organisms, with a conserved set of enzymatic activities and chemical reactions and a wide range of structural features.

Despite the conservation of the biosynthetic reaction mechanisms, the architecture of FAS differs across kingdoms of life. Its structural organization ranges from free-standing enzymes that cooperate creating an assembly line (Type II) [47], to the cross-shaped homodimer with seven different active sites per dimer found in higher eukaryotes (Type I) [48,49], to the barrel-shaped fungal FAS megasynthase (Type I) with its 48 functional domains embedded in a matrix of scaffolding elements that comprises almost 50% of the total sequence [50–52]. This paradigm of evolution has led to more questions than answers, but its structural variability might be a gift for enzymatically engineering the FAS system.

De novo fatty acid biosynthesis in animals is carried out iteratively across the Type I FAS. Its primer substrate, acetyl-CoA, is repeatedly extended with malonyl building blocks, two carbons at a time, releasing CO_2 upon condensation. The acyl carrier protein (ACP) substrate shuttling domain is responsible for transporting the substrate from one domain to another. The product is released once the chain-terminating

thioesterase (TE) domain cleaves the thioester bond between the product and ACP's prosthetic arm. Animal FAS megasynthases have a narrow range of products: longer fatty acids, namely palmitic (16C) or stearic acids (18C). TE is the domain responsible for determining the length of the product [48,53]. The mechanics of such specificity are yet to be understood [54].

The lack of X-Ray crystallographic structures of the mammalian FAS with all of its active domains resolved has hindered the study of this multidomain enzyme. Grappling with this shortcoming in 2018, our group has established the first complete structure of the mammalian Type I FAS employing homology modeling and molecular dynamics methods. Furthermore, we have studied and established the catalytic mechanism of all of its catalytically active domains except TE and ACP. Nevertheless, we devised a QM/MM model for the future study of the mammalian FAS TE mechanism, found in subchapter 3.3.

S. cerevisiae has been the preferred target of many enzymatic engineering studies. Like the mammalian FAS, the yeast FAS Type I, a multidomain megasynthase with its structural organization resembling a barrel. However, comparative analysis between animal, fungal, and bacterial FAS showed that Type I are differently organized, with the most striking difference residing on the mechanisms of substrate shuttling during fatty acid *de novo* biosynthesis.

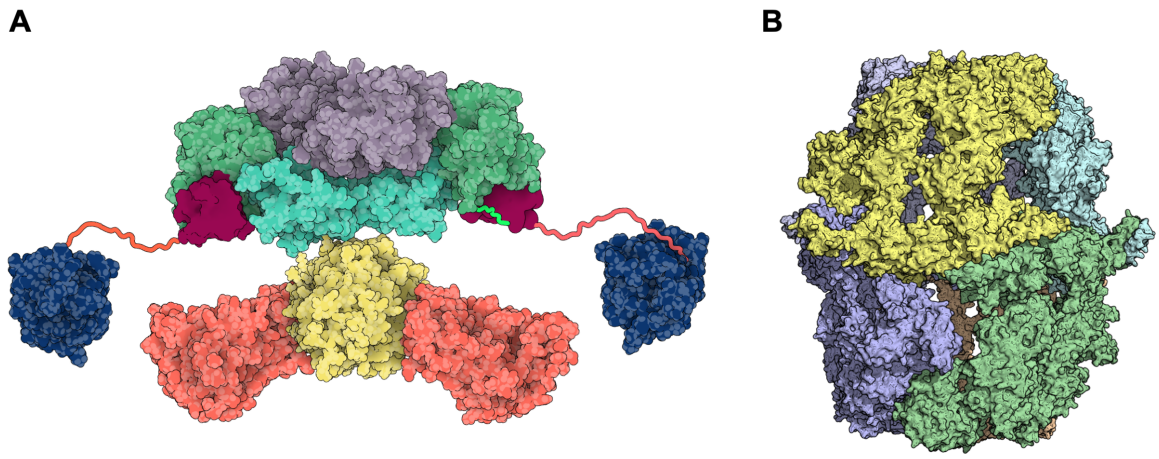


Figure 1| Representation of the Type I FAS. A) Representation of the animal FAS homodimer head-to-head model, as obtained from homology modeling [55]. B) a representation of the dodecameric assembly of the barrel-shaped fungal FAS

The first crystal structure of yeast FAS was obtained in 2006, from the fungus *Thermomyces lanuginosus*, at 5 Å resolution [56]. In the following years, additional X-ray crystallographic structures of yeast FAS have been determined with improved resolution [51,52]. *S. cerevisiae* FAS megasynthase comprises 48 active domains, and it is the most kinetically efficient out of all FASs systems [57,58]. Therefore, it has been hypothesized that its individual FAS catalytic units have merged throughout evolution to form a single multi-functional polypeptide FAS, expressed from a single gene.

As for Type II FAS systems, seen in bacteria, plants, and mitochondria [59], its intervening enzymes are free-standing and individual, forming a metabolic pathway to produce fatty acids. Considering the diversity of organisms Type II FAS are found in, there is a wide range of condensing enzymes involved in the fatty acid biosynthesis, such as the β-ketoacyl ACP synthases I (KAS I or FabB), II (KAS II or FabF), and III (KAS III or FabH) [47,60–64] with a variety of substrate specificities. As such, the wide range of substrate specificities among these condensing enzymes gives rise to a panoply of fatty acid products formed in the respective biosynthetic pathways.

FAS systems are remarkably robust and elastic, allowing for mutation, deletion, and addition of domains to the original scaffold while maintaining functionality. This was first explored using the mutant complementation approach, where different domains exhibited knockout mutations, to understand how ACP iteratively interacted with the neighboring catalytic domain – whether this interaction was limited to domains in the same dimer, or opposed dimer, or both [65]. FAS Type I functionality elasticity was further pushed through the heterologous integration of bacterial enzymes, particularly ester hydrolases, into the central wheel of yeast FAS has been a successful approach to alter fatty acid chain specificity [66,67]. For example, the heterologous expression in *S. cerevisiae* of a fungal FAS from *Rhodosporidium toruloides,* where a second ACP domain was replaced by a short-chain specific bacterial TE domain from *Acinetobacter baylyi* [68], resulted in a 20-fold increase in the production of hexanoic and octanoic acids compared to the wild-type form [66].

Favorable results employing engineered FAS for biodiesel production include the production of FAEE for the production of biofuels in *S. cerevisiae* [18]*,* and the fatty alcohol 1-octanol, which has been described as a valuable substitute for diesel and jet fuels. Previous research has devised an extensive portfolio of approved molecules from microbial production [69–72] where 1-octanol stood out as a substitute for diesel and jet fuels [73–76]. By comparing several characteristics of fossil-derived and bio-derived diesel fuels with saturated short- and medium-chain alcohols, it was shown that 1-octanol exhibits best matching overall properties compared to ethanol or other long-chain alcohols [76]. Octanol-1 has been produced in small amounts from fermentable sugars employing engineered bacteria via reduction of thioesterase-mediated premature release of octanoic acid from fatty acid synthase, or via a reversal of the β-oxidation pathway [77].

1.3.2. Mesophilic *vs* extremophilic enzymes in industrial processes

Microorganisms are ubiquitous forms of life. They exist and survive in different conditions, and can be classified according to the environmental surroundings they

thrive in. This classification originated from the anthropocentric perspective that the conditions paramount for human colonization (1atm pressure, neutral pH, salinity, temperatures between 20 to 40 °C) are deemed standard, and all else considered "extreme." Microorganisms that inhabit such "extreme" conditions are called extremophiles. They thrive in a variety of environments leading to a categorization based on their optimal growth conditions, both physical as temperature (−2 to 12 °C, 60 to 110 °C), pressure (0 to 110 MPa) or radiation (high, supra-lethal doses of UVR, X-Ray, gamma-rays [78]), and geochemical such as salinity (2–5 NaCl) and pH (<2, >9) [79].

Extremozymes, extremophiles' enzymatic machinery, offer advantages for industrial processes due to their unrivaled performance latitude. Therefore, extremozymes' structural features have been studied and investigated to decode the mysterious cocktail that allows these extremozymes to thrive in human-deleterious conditions.
The research community has tried to pin down the many characteristics and structural features that contribute to extremozymes' resilience to external environmental conditions. Observing which organisms survive and thrive in increasingly extreme conditions offers an indication of which factors might be at play.

1.3.3. Extremozymes structural properties that make up their armor

To achieve thermal stability and denaturation resistance, there are many factors at play - a higher number of hydrophobic residues at the core of the protein and ionic interactions, a higher ratio of disulfide bridges [80] and proline residues [81], all of which contribute to greater packing and rigidity. Comparative studies have narrowed down to four the number of significant determinants of structural thermostability of thermophilic enzymes.

Ion pairs are oppositely- or like-charged groups found at ≤4 Å, whose primary role is to maintain the folding and stability of the enzyme's tertiary structure. In 1983, Barlow et al. studied the role of ion-pairs in 38 globular proteins: an average of 33% of the protein's charged residues were involved in ion-pairs, 76% of which contribute to the

stabilization of the tertiary structure [82]. A more recent study focused on the hyperthermophilic ribosomal protein L7Ae of the *Aeropyrum pernix* (110 °C melting temperature) has shown an extensive ion-pair network in the hyperthermophilic protein when compared to the human homolog (19 out of 24 ion pairs are involved in the network *vs.* zero ion-pairs in the human counterpart) [83].

Hydrogen bonding in the context of thermophilic enzymes often displays a distinctive pattern of exposed residues establishing hydrogen bonds with buried residues creating a crosslinked network that contributes to the protein's stabilization in denaturing conditions [84].

Another contributing factor to the thermophilic enzymes' resilience is the hydrophobic environment found at their core. A study has shown that 80% of analyzed thermophilic enzymes displayed higher hydrophobicity at their core than their mesophilic equivalents [85]. **Hydrophobicity of buried side chains** was shown to be the key driver behind the increased compaction level found at the core of studied thermophilic proteins – as these hydrophobic residues interact and adhere to one another, they create a reinforced concrete-like scaffold at the core of the enzyme, counteracting external destabilizing influences that could jeopardize the protein's integrity. This was hypothesized and corroborated in a number of studies [86,87]; one study compared the mesophilic and thermophilic equivalents of methanococcal adenylate kinases whose melting points differed by 34 °C. Despite the 68% sequence identity between the two, a small number of non-conserved hydrophobic residues were found in the core area of the thermophilic methanococcal adenylate kinases at close proximity [88].

Compact tertiary structure cores, resulting from a higher percentage of small amino acids in detriment of bulkier ones [89], a wide distribution of proline residues whose pyrrolidine ring creates steric hindrance in the surrounding area [90–92], alongside a higher number of disulfide bonds [93,94], all contribute to a higher resistance to unfolding and tighter packing [95–97].

Altogether, these structural features generate increased resistance to extreme conditions, allowing thermophilic enzymes to survive in conditions where their mesophilic counterparts would perish, thus making thermophilic enzymes, and extremozymes in general, appealing for industrial applications.

1.3.4. Enzymes for biomass processing

Lignocellulose is the most abundant biopolymer, containing mostly polysaccharide sugars which can be used for many applications, including as sugar monomers for biofuel production. Lignocellulose abundance as waste biomass makes it an undeniable energy source waiting to be fully explored. It can be an alternative, renewable source of energy, in opposition to fossil fuels, helping to relieve global climate change.
Lignocellulose's main components are carbohydrate polymers - cellulose and hemicellulose (which account for 30-50% and 15-30% of the dry weight of lignocellulose, respectively), and lignin accounting for 25-40% of plant dry mass [98–101].

To use lignocellulosic biomass's potential, an extensive protocol of treatments and processing follows. One of its many hurdles is the complex matrix of lignin-carbohydrate-complexes (LCC) [102]. LCCs are covalent linkages between lignin and hemicellulose in the plant cell wall, and they're partly responsible for the recalcitrance of lignocellulosic matter – a phenomenon that hinders separation of cellulose, hemicellulose, and lignin, affecting the use of plant biomass as prime-matter in industrial context [1,102]. Furthermore, current methods of LCC processing employ severe pretreatments, consuming both energy and polluting chemical reactants[103,104]. These disadvantages call for cleaner alternatives, namely the consideration of enzymes to catalyze LCC processing efficiently. There are three main types of LCCs: esters, benzyl ethers, and phenyl glycosides [1,102,105]. Depending on the linkage type, different kinds of enzymes are apt for the job. For example, amongst the ester group, linkages between (4-O-methyl)-glucuronic acid residues in xylans and lignin alcohols are found in woods and grasses and can be degraded by Glucuronoyl esterases (GE)

[106]. Furthermore, the search for carbohydrate degrading enzymes in extremophiles organisms has signaled other enzymes eligible for LCC processing in industrial conditions. For instance, a notable application of these carbohydrate extremozymes is their use in degrading recalcitrant lignocellulosic biomass in biorefineries.

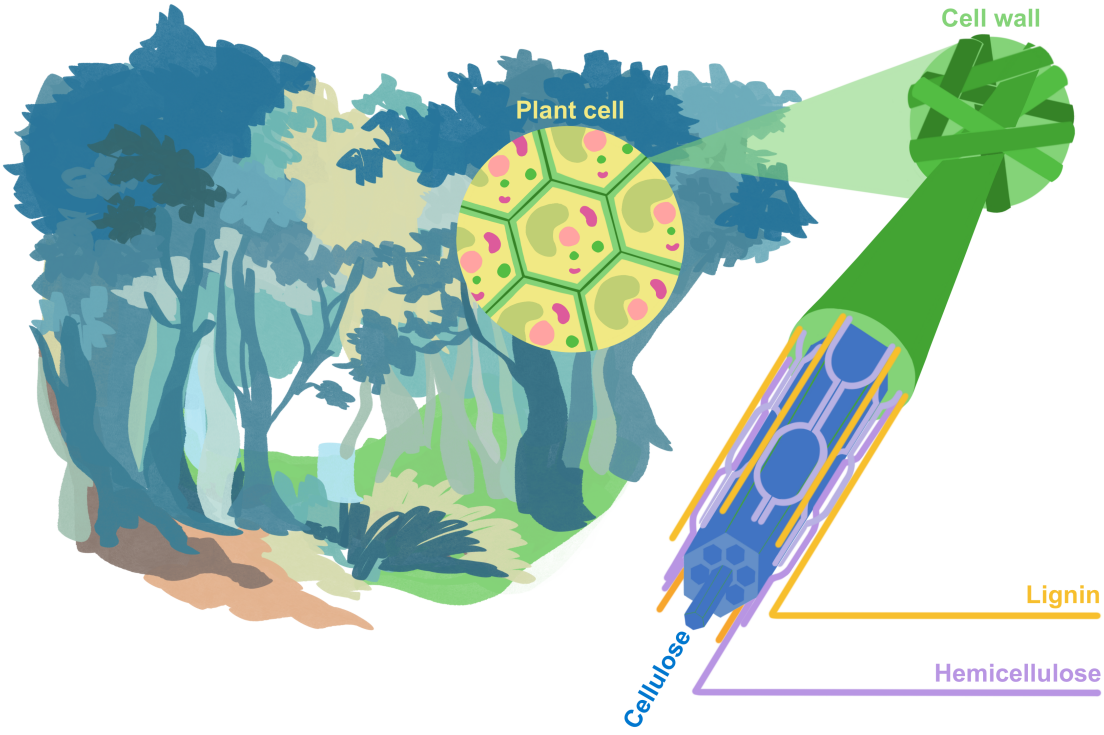


Figure 2| Lignocellulosic biomass. Illustrative and schematic representation of lignocellulosic biomass structure composed by cellulose, hemicellulose and lignin, where the covalent linkages of lignin-carbohydrate complexes (LCC) are found. Biocatalysts for the processing of LCC contribute to the reduction of costs and pollution. Illustration © Mafalda Salgueiro.

1.3.5. Thermophilic Glucuronoyl Esterase

GEs are involved in microbial degradation of lignocellulosic plant matter, cleaving LCCs. The first carbohydrate esterase was discovered in 2006 by Spanikova [107] in a wood-rotting fungus called *Schizophyllum commune,* which presented cellulolytic activity. Enzymatic assays showed its ability for hydrolyzing unnatural, synthetic esters of 4-*O*-methyl-*D*-glucuronic acid and glucuronic acid, thus being named GE [107]. In 2007, the gene encoding GE was identified in *Trichoderma reesei* (anamorph of

Hypocrea Jecorina). A subsequent gene blast revealed the presence of GE coding genes in several microbial genomes. Its occurrence in different organisms led to introducing a new carbohydrate esterase (CE) family called CE15 in the Carbohydrate-Active Enzymes (CAZy) classification [108]. Members of the CE15 family are found in fungal and bacterial organisms, all of which exhibit GE activity, cleaving esters of D-glucuronic acid. Interestingly, several bacterial CE15 enzymes were also found to be active on esters of galacturonoate, besides *D*-glucuronic acid [109], suggesting a broader substrate specificity.

CE15 enzymes, including GEs, belong to the α/β-hydrolase superfamily, which c a catalytic triad common with serine-type hydrolases: Glu/Asp-His-Ser [109–112]. All CE15 enzymes exhibit the canonical catalytic triad except a particular group of the CE15 family that displays a serine residue in place of the glutamic acid at the catalytic triad [113].

1.4. Enzymatic engineering and yield optimization methods

The broad subject of this thesis is how can enzymes drive our society to a more ecologically-oriented functioning? How is the scientific community working towards the goal? The driving force behind most societal development is profit, and the ultimate goal of applied chemistry is to produce useful molecules efficiently and inexpensively. Enzymes and their ability to take chemical building blocks and transform them into valuable products in a kinetically favorable process have been regarded as the unattainable holy grail of industrial catalysis. Enzymatic engineering is striving to harness the catalytic power of enzymes and use it in industrial processes.

The amino acid sequence of an enzyme determines both its function and its three-dimensional structure. Using this as the starting point for enzymatic engineering, scientists can use distinct approaches to improve an enzyme's reaction rate or alter its catalysis altogether, with varying degrees of success as mutations will most often destabilize the protein's integrity. Thus, successful enzymatic engineering depends on the compromise between protein stability and desired functionality [114].

Current methodologies and design methods for enzymatic engineering can be divided into semi-rational and rational approaches, for example, directed evolution or site-directed mutagenesis based on prior knowledge [115,116].

Directed evolution is a **semi-rational design** method used to mimic the process of natural evolution on a reduced time scale. This requires several rounds of mutagenesis and screening, starting from a selected enzyme. For example, random mutants can be created from a parental protein through an iterative process employing DNA shuffling [117–119], error-prone mutagenesis [120,121], chemical mutagenesis [122,123], site-saturation mutagenesis [124,125], or different mutator strains [126,127]. However, these methods are unpredictable and require extensive analysis and library processing. Therefore, identifying and isolating mutants with the desired features and properties is the bottleneck of directed evolution. Upon creating a library of mutants, managing the library can be achieved through two main methods: screening and selection. In the selection method, the quality of the mutants is assessed based on a direct correlation between cell growth and an optimized enzyme function. In the screening method, all mutated variants are tested for the desired feature, assessing its qualities individually, and minimizing potential false negatives. Once the best performing mutants are identified, there's the possibility for a new round of directed evolution, repeating the screening and selection of protein mutants until the desired feature or property is found. It's been generally believed that the best starting point or a directed evolution campaign is a protein whose sequence is either based on a protein family of interest or closer to the protein ancestor, assuring the most suited scaffold for directed evolution and its mutation-associated pressure. A successful case of a directed evolution campaign is the improvement in the thermostability of galacto-N-biose/lacto-N-biose I phosphorylase from *Bifidobacterium longum* JM1217 [128].

Besides directed evolution, a semi-rational method, the modification of the properties of an enzyme could be accomplished through rational approaches. Rational design methods rely on computational design to predict stabilizing mutations in the context of a given fold to provide maximum catalytic efficiency while minimizing disruption of the structure of the active site.

Rational design methods are based on extensive knowledge and empirical data about the enzyme's structure and function. It resorts to multiple methods, which include a comparison to homologous proteins [129], phylogenetic analysis [130], optimization of residues and loops showing unfavorable Ramachandran angles and high B-factors, optimization of charged interactions (salt bridges and hydrogen bonds) [131], structure-based computational design[132,133] and methods based on the calculation of free energies [134]. As a consort, these methods predict potentially destabilizing mutations that would hinder the protein's folding or active site architecture while potentiating other stabilizing interactions.

To achieve the thermostability of a given enzyme, one route is to search for its thermophilic homolog and apply it to industrial settings. However, the process isn't always as straightforward. In the absence of a thermophilic homolog, the alternative process involves three steps: selection of an appropriate enzyme, identification of residues to be mutated, and characterization of those mutants. Once characterized and identified the mutations that potentially improve the enzyme's stability, the final step is to express the mutated enzyme in a heterologous host such as *S. cerevisiae*. As rational strategies rely on structure-based studies, having an atomistic level of understanding of the catalytic mechanism of enzymes is key to the development of improved mutants [135]. Besides empirical wet lab studies, knowledge about enzymes can be obtained from public databases such as PDB [43,136].

Extremozymes mechanisms for resistance to external factors are both advantageous and a caveat for enzymatic engineering due to its many playing factors. Thus, the implementation of rational and semi-rational design strategies allows for an extensive survey of mutants that will eventually result in engineered enzymatic catalysis suited for efficient industrial catalysis.

1.5. Enzymes as therapeutics: the case of snake venom

In previous subchapters, the importance of enzymes in day-to-day human activities (such as baking and brewing) and for industrial applications was made clear. But enzymes and proteins may also play an essential role in therapeutics. For the past 100 years, several types of therapeutic enzymes, such as L-asparaginase, streptokinase, collagenase, ribonuclease, uricase, glucosidase, etc., have gained important applications in the pharmaceutical industries [137].

One bottleneck for using enzymes as therapeutic agents is the discovery process – finding organisms that produce enzymes with desirable properties and identifying such enzymes. The study of snakes (ophiology) and biological toxins (toxinology, not to be confused with toxicology) have demonstrated the potential of snake venom as a reservoir for enzymes and proteins with medicinal properties. Snake venom has been used as an ailment treatment for thousands of years, especially in traditional Chinese medicine [138]. Nowadays, medicines like Captopril® (Enalapril), Integrilin® (Eptifibatide), and Aggrastat® (Tirofiban) are FDA (Food and Drug Administration) approved drugs based on snake venoms, and they are commonly prescribed. Snake venom offers an extensive catalog of molecules that range from 100 to 500 pharmacologically active compounds. However, no more than 0.01% of these compounds have been identified and characterized, and a large proportion of toxins act on unknown receptors. Moreover, most known toxins have been described only partially, with catalytic mechanisms still unknown, such as the secreted Phospholipase A_2 ($sPLA_2$).

$sPLA_2$ is omnipresent in all organisms, exists in all human cells, and is thought to partake in the regulation of phospholipid metabolism [139]. PLA_2, and $sPLA_2$, are responsible for phospholipid hydrolysis, releasing fatty acids and lysophospholipids. In the context of snakebite, $sPLA_2$s cause the disruption of cellular membranes, resulting in cell lysis. But studies have shown that $sPLA_2$ also possesses antibacterial activity [140]. This is a trend observed in many other proteins and enzymes found in

snake venom: though deleterious to the bitten prey, these enzymes, when isolated and studied, reveal enormous potential as therapeutic agents.

1.6. Science communication and outreach

For centuries, science was limited to upper-class individuals with access to higher education. But with the advancement of technology, scientific literacy is essential for adequate participation in our society. As a result, there are two types of science communication: *i)* science outreach or outward-facing, in which professional scientists communicate with people outside the scientific community; and *ii)* science "inreach," or inner-facing, established between experts of a similar academic level [141].

Communicating science and technology is a priority for many research institutions – it allows the dissemination of knowledge to a general audience, garners respect from policymakers, entices new students towards science education, and is a cornerstone of securing funding. It is also a viable path for establishing new collaborations with like-minded scientists and promoting research, especially if pledging to the Open Access, the principle and practice of providing free online access to scholarly publications [142].

In the context of the four years of doctoral studies, efforts have been made to contribute to both science outreach and "inreach." Two articles were written for the Portuguese general public, explaining the importance of biofuels for a more environmentally conscious society [143], and a survey of COVID-19 viral proteins available on PDB [144]. For students of Bioinformatics and Computational Chemistry, a workbook was devised for the Bioinformatics course taught at the Faculty of Sciences of the University of Porto, and a QM/MM protocol taught at the 2019 Dynapeutics International Summer School in San Sebastián, Basque Country (Spain).

1.7. Outline of the thesis

In the previous subchapters, a case was made on the climate crisis we are facing and the crucial role of biofuels and green chemistry as tools to diminish greenhouse gas emissions and the toll of pollution. Enzymes have been at the forefront of green chemistry as the highest standard for environmentally-friendly catalysis. Yet, despite their immense power and potential, enzymes stand as undomesticated entities, as humans still struggle to make use of their abilities in their favor.

The use of computational chemistry in the study of enzymatic catalysis made strides in the atomistic understanding of enzymes and their catalytic mechanisms. In this thesis, we applied QM/MM methodologies to analyze enzymes with potential for biofuel production.

Chapter 3 launches with an introduction to the realm of fatty acid synthesis, its catalytic mechanisms, and its potential for therapeutics and industrial applications, based on excerpts of the review “Animal Fatty Acid Synthase: A Chemical Nanofactory.” This work is followed by the first complete human Type I FAS model with its fourteen active domains. The model was solved using homology modeling and protein-protein docking techniques introduced in the Methods section. Next, the study of the Type I FAS is further explored with the modeling of the TE domain, responsible for the cleavage and release of the synthesized fatty acids, and proposal of a QM/MM model of the TE-Palmitoyl complex for future catalytic mechanism studies using QM/MM ONIOM methods. Finally, to further explore the potential of the megasynthase system for the production of desired molecules, an extensive review on the multi-domain PKS is introduced. PKS systems, similarly to Type I FAS, are megasynthases, and they are responsible for the production of polyketides in bacteria, fungi, plants, and a few animal lineages. The review “Animal Fatty Acid Synthase: A Chemical Nanofactory” is available in the Appendix section.

Chapter 4 delves into thermophilic enzymes with an introductory review on thermophilic enzymes and their application in greener industries “Thermophilic

Enzymes: The Key for Greener Industrial Processes," followed by the study of *Tt*GE, a Glucuronoyl esterase enzyme with thermoresistant properties whose catalytic activity contributes to the degradation of lignocellulosic biomass – needed for biomass processing industries and biofuel production. *Tt*GE catalytic mechanism was studied in detail using QM/MM ONIOM methods, resulting in the characterization of the ten stationary points of its reaction mechanism, alongside potential mutations that could be further explored to improve the catalytic rate of the enzyme.

Throughout this thesis, collaborations with other groups followed. The scope of this thesis was extended to the study of snake venom and, in particular, of the secreted Phospholipase A2 enzyme found in snake venom. **Chapter 5** encompasses the "Structural, enzymatic and pharmacological profile of AplTX-II - A basic $sPLA_2$ (D49) isolated from the *Agkistrodon piscivorus leucostoma* snake venom". For this study, we have resolved the structure of AplTX-II with the docked synthetic substrate NOBA (4-nitro-3-octanoyloxy benzoic acid) by homology modeling and performed MD simulations with explicit solvent.

Science communication was also a key aspect during the four years of the Doctoral Program in Chemistry, emphasizing science education and outreach. **Chapter 6** encompasses two articles published in the magazine Ciência Elementar, aimed at the Portuguese general public and high-school students, alongside a QM/MM protocol taught at the 2019 Dynapeutics International Summer School in San Sebastián, Basque Country (Spain). On the same subject of science communication, a workbook for the Bioinformatic course taught at the Faculty of Sciences of the University of Porto is available in the Appendix section.

Finally, in **Chapter 7**, a unifying conclusion is offered to end the present thesis. In **Chapter 8**, an additional paper titled "Evaluation of a flavonoids library for inhibition of pancreatic α-amylase towards a structure-activity relationship" is presented, with subject matters falling outside the scope of the present thesis but nonetheless under the umbrella of computational methodologies.

The next chapter will offer the theoretical background necessary to understand the methods employed in the chapters mentioned above.

CHAPTER 2. Theory behind computational methods for the study of enzyme catalysis

Upon gaining awareness of the power of enzymes for all kinds of human-related activities, from medicinal to industrial, the scientific community has tried to tackle the understanding of enzymes' functioning at a molecular and atomistic level. Enzyme-catalyzed chemical reactions involve bond breaking, and bond making, at a fast time scale, usually in the femtoseconds to nanoseconds order. At the center of studying these biomolecules, Biochemistry employs experimental methods such as spectroscopic techniques, mutagenesis experiments, and kinetic evaluations. It offers insights into conformation, structure, atoms involved, and kinetics of the reactions. However, none of these disclosed the origin of the rate acceleration ability of enzymes. Moreover, as the half-life timescale of its transition state structures is often in the sub-picosecond, it is unfeasible to be experimentally detected.

Since the dawn of theoretical chemistry in the 20th century, its primary goal has been to understand, predict and reproduce properties at the atomistic level. Schrödinger's equation offered a glimpse into the behavior of atoms, but its complexity prevented calculations from going beyond that of the hydrogen atom. With the advent of computers, growing computational power, and the devise of more efficient algorithms, theoretical and computational methods are now at the very foundation of molecular modeling. This field refers to all methods used to describe and simulate, *in* silico, molecular systems such as enzymes, with atomistic detail [145,146]. Current computational methods for studying enzymatic reactions can be divided into two main levels of theory: Molecular Mechanics (MM) and Quantum Mechanics (QM).

When studying the phenomena in the nanosecond to microsecond timescale, MM methods are the preferred route. Despite not being adequate to study and describe catalysis, the main advantage of MM stands with its ability to perform extensive conformational sampling useful for structural and thermodynamic studies [147–149].

When the subject of the study is found in the femtosecond to nanosecond order, such as catalytic reaction and its transition states, QM or hybrid QM/MM methods are best suited. A hybrid approach allows for a compromise between computational cost and an accurate description of the system [145,150,151]. Most importantly, when choosing a method, it is crucial to assess the existence of electronic rearrangement.

In the present thesis, we will refer to methodologies employed in studying enzymatic catalysis, focusing on Homology Modelling, Protein-Protein Docking, followed by Molecular Dynamics (MD) and QM/MM methods.

2.1. Homology Modelling

PDB is responsible for hosting over 180 hundred experimentally resolved three-dimensional (3D) structures [43,136], determined using methods from X-Ray diffraction [152] to nuclear magnetic resonance (NMR) [153], to cryogenic electronic microscopy [154]. However, the number of annotated protein sequences, including redundant structures, surpasses 215 million, according to the 2021 announcement by the National Center for Biotechnology Information (NCBI) [155]. The disparity between the number of identified protein sequences and the number of resolved 3D structures (number of identified protein sequences is 1159 times larger than the number of protein 3D structures) is proof of the need for methods that can make up for the lack of experimentally determined 3D structures of proteins. As an alternative to experimental methods, computational methods allow predicting 3D structures of proteins based on their amino acid sequence [156].

Homology modeling is the most accurate of methods [157,158], as well as the most accessible to predict the 3D structure of proteins based on its amino acid sequence [159]. The protocol for homology modeling can be briefly summarized as follows:

1. Identification and selection of templates based on sequence similarity to the target sequence in PDB [43,136] or similar databases, usually using the Basic Local Alignment Search Tool (BLAST) [160] for pairwise sequence-sequence alignment or other similar algorithms. Besides sequence similarity, which

should be above 25%, phylogenetic similarity, common environmental features factor in the process of template selection, and resolution of the eligible template must be considered to ensure a good quality model.

2. Sequence alignment and alignment correction through careful inspection using Clustal W. or similar alignment tools. Alignment errors by a single residue cause shifting of α carbon, affecting following residues conformation.
3. Model building can be achieved through different methods. These are classified as segment matching methods [161], rigid body-assembly methods [162], spatial restraint methods [163], and artificial evolution methods [164]. Our method of choice was rigid-body assembly in which the protein structure is segmented into conserved core regions, loops, and side chains. Then, by dissecting the protein structure, a final protein 3D structure can be assembled by bringing together the rigid bodies that were modeled from the aligned template protein structure. A well-known tool for this method is the free, online webserver SWISS-MODEL [159].
4. Loop modeling is crucial. As loops show increased structural variability compared to strands and helices, prediction of their structure is more complex, thus loop regions are more prone to modeling errors [165].
5. Model optimization through energy minimization calculations using MM force fields [166]. More on the topic in subchapter 2.3.
6. Analysis of stereochemistry parameters is necessary to ensure the quality of the model. This is considered an internal evaluation where features such as bond length, torsion angles, rotational angles, disulfide bonds, and others are inspected to evaluate the model's overall quality. Algorithms such as PROCHECK [167], Molprobity [168] and Ramachandran Plot are popular tools to detect "atypical" regions in the model that may have originated during the modeling process and need to be further inspected.
7. Analysis of the homology model folding using algorithms, such as ProsaII [169] or QMEAN [170] (a tool integrated into SWISS-Model [159]), attributes an estimated quality score to the model based on the geometrical analysis of a pool of single models.

8. Validation of the model through Molecular Dynamics simulation, ensuring optimization of loop region and side chains, and refinement of the generated models.

Homology modeling has become instrumental in the drug discovery process, particularly when high-quality models of the target protein are unavailable. An interesting case that illustrates this is that of the novel acetohydroxy acid synthase (AHAS, EC 2.2.1.6.) inhibitors against *M. tuberculosis.* At the moment, the 3D structure of *M. tuberculosis* AHAS remains unresolved. Thus, homology modeling allowed creating of a 3D structure of the target enzyme, which was then used as the target protein in docking simulations. From the docking simulation, 23 novel AHAS inhibitors were identified, three exhibited more activity than other compounds with antituberculosis activity [171]. Thus, homology modeling is an exemplary tool for molecular modeling unresolved 3D structures of enzymes.

2.2. Protein-Protein Docking

Protein-protein complexes are challenging to crystalize, creating a deficit of available structures on PDB. Lack of available structures hinders the understanding of protein-protein interactions, which negatively impacts pharmaceutical and medicinal sciences, especially in the design of compounds to affect protein-protein interaction.

As explained in the previous subchapter, computational methods' growing accuracy and reliability have compensated for the difficulties of X-ray crystallography and NMR in dealing with transient complex formation and high molecular weight, respectively [172,173]. Computational methods used to study and predict protein-protein complexes are termed protein-protein docking. These docking procedures rely on various theoretical algorithms to assemble two separate protein components into a biologically relevant and plausible complex structure, useful for potential applications in rational drug design and protein engineering.

Currently, the only method that can systematically give insights into the many protein complexes encountered in biological processes is *in silico* protein docking. Docking methods are based on the idea of complementarity between the interacting molecules. The objective of protein-protein docking is to predict the three-dimensional arrangement of a protein-protein complex from the coordinates of its free-form (unbound) experimental or modeled individual three-dimensional structures, generating accurate predictions that will point out most of the residue-residue contacts involved in the target interaction.

Several software packages have been developed for *in silico* protein-protein docking assays. The majority offer prediction of protein-protein complexes using either geometric or electrostatic factors, or both. The search for those contacts requires an exhaustive exploration of one protein's rotational and translational space to the other, resulting in a six-dimensional one search.

For the study of FAS and the many points of ACP transient interaction with various catalytic domains, we used the tool HADDOCK. HADDOCK stands out from other approaches for it is based on pre-existing experimental data, bioinformatics information, and mutagenesis data to efficiently drive the docking process [173,174]. In the absence of any experimental information, bioinformatics interface predictions can be used. They are typically based on a comparison of the sequence and structural features of the protein target against information contained in databases to predict residues located at the interface. Several web servers are available for this purpose, some of which have been integrated into the interface meta-predictor CPORT especially developed for use with HADDOCK [175]. All gathered data on the interacting residues is introduced as ambiguous interaction restraints (AIRs) to drive the docking in HADDOCK.

Once the HADDOCK run finishes, the generated structures are automatically ranked according to their intermolecular energy, and clustered using a “7.5 Å cutoff based on their pairwise ligand interface RMSd (Root Mean Square deviation) values, and the

cluster ranks are determined according to the average score of the four best structures of each cluster" [176].

Visual inspection of the generated poses is strongly encouraged, by analyzing distances between interacting residues on opposing entities of the protein-protein complex, and by further validation through MD simulation. This protocol is used and further explained in the published work "Modeling of Human Fatty Acid Synthase and in Silico Docking of Acyl Carrier Protein Domain and Its Partner Catalytic Domains" [55].

2.3. Molecular Mechanics

As explained briefly above, MM methods can be used to study biochemical molecules at the nanosecond to microsecond timescale when no significant electronic rearrangements are present. Their simplistic take is based on classical Newtonian mechanics, *i. e.*, Newton's laws of motion, which describes molecules as an agglomerate of spherical particles coupled by harmonic strings, and where the movement of each particle is defined by Newton's second law:

$$F = ma$$

Equation 1

F is force on a particle, m is mass of the particle, and a is the particle's acceleration.
In MM methods, each atom is simulated as a "ball" (particle) with a set of parameters, such as its van der Waals radius and atomic charge, and the bonded interactions between "balls" are treated as harmonic springs [148,149].
Thus, the system's total energy can be calculated based on the atomic coordinates. The inclusion of electrons and all of the quantic aspects of the nuclei movement are achieved through parameterization of force fields. This approach allows for the study of systems with a significant number of atoms - for instance, the study of protein conformational flexibility - without the computational costs of QM calculations. There are many different force fields described in the literature, each aimed at a particular set of molecules.

Most often, MM methods are not suited for the description of chemical reactions because these require breaking and bond making and depend on electronic distribution. For this type of study, QM methods are best-suited. Despite the growing availability of QM methods and increased computational power, the inherent computational cost to QM level calculations limits the size of the system and the simulation scales. This severe limitation often excludes QM methods from considering the dynamic evolution of a system, thus hindering our theoretical understanding of key factors affecting the overall behavior of a system. To counter this issue, empirical force fields are being trained using QM energy and structure data. These force fields require significantly fewer computational resources, therefore supporting simulations to better describe dynamic processes. Such empirical methods, including reactive force-field (ReaxFF) [177] trade accuracy for lower computational expense, make it possible to reach simulation scales that are orders of magnitude beyond what is tractable for QM. A recent study combining MD simulations and ReaXFF was able to simulate the microscopic reactive behaviors of brown coal combustion in the O_2/CO_2 atmosphere by thoroughly tracing the motion trajectories of atoms, as well as the formation, transition, and breaking of bonds between atoms, the generation pathways of CO_2 and the effects of temperature (1600–2000 K) and oxygen concentration (21%~30%) on the coal oxy-fuel combustion [178].

2.3.1. Force Field

In MM methods the system's potential energy is calculated using force fields that parcel the potential energy into terms. The most generic biomolecular empirical force field describes the potential energy (V) as the sum on bonded (covalent) and nonbonded (non-covalent) energy terms:

$$V = \sum_{bonds} \frac{1}{2} k_r \left(r - r_{eq}\right)^2 + \sum_{angle} \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} K_\varphi [1 + \cos(n\varphi - \gamma)]$$
$$+ \sum_{\substack{improper \\ torsions}} \frac{1}{2} k_\varphi (\varphi - \varphi_0)^2 + \sum_{\substack{nonbonded \\ pairs\ ij}} \frac{q_i q_j}{4\pi r_{ij}}$$
$$+ \sum_{\substack{nonbonded \\ pairs\ ij}} \varepsilon_{ij} \left[\left(\frac{R_{min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min,ij}}{r_{ij}} \right)^{6} \right]$$

Equation 2

The first four terms describe the bonded energy terms. The first term describes the bond stretching energy between covalently bonded atoms, the second term refers to the angle bending energy between three bonded atoms, and the third and four refer to dihedrals or torsional terms and improper torsions.

As for the nonbonded energy terms (the last two), they refer to the non-covalent bonding energy between all atom pairs, and are decomposed into van der Waals and electrostatic energies.

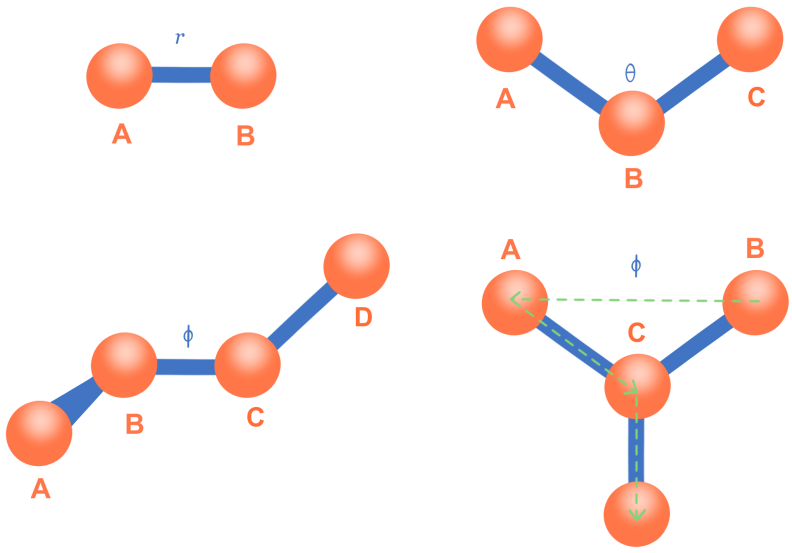


Figure 3| Schematic representation of the four bonded energy terms. r represents the bond length between atoms A and B; θ is the bond angle between atoms A, B, and C; φ is the dihedral torsion between atoms A, B, C and D, and the φ is the angle formed by two non-concurrent vectors in the plan. Scheme © Mafalda Salgueiro.

Bond stretching energy is the potential energy associated with the elongation or shrinkage of a bond between two atoms, being this bond simple, double, tiple, or aromatic. The force field makes an approximation of this bonded interaction as a harmonic potential.

$$E_{bonds} = \frac{1}{2} k_r \left(r - r_{eq}\right)^2$$

Equation 3

In this equation, r is the present bond length and r_{eq} is the equilibrium bond length, and k_r is the force constant, that describes the harmonic regime near the equilibrium. As bond breaking or making is usually out of the scope of MM methodologies, the quadratic approximation is suited to MM methods as it aptly describes near minimum conformation. However, when the internuclear distance is at the farthest extremes, the description is no longer adequate. This is the case of: *i)* when the distance between the two atoms decreases, and the potential energy quickly increases due to the Pauli repulsion; or *ii)* when the distance between the two atoms surpasses the equilibrium bond length, and the potential energy increases quickly go then slowly converge to the point where there is no interaction between atoms. In both cases Morse potential energy function describes the potential energy more accurately than the quadratic approximation.

Angle bending energy is the potential energy associated with the angle between three bonded atoms and, similarly to bond stretching energy, it is described by a quadratic equation:

$$E_{angle} = \frac{1}{2} k_\theta (\theta - \theta_0)^2$$

Equation 4

Where k_θ is the force constant, θ_0 is the equilibrium angle and θ is the present angle. The quadratic potential for angle bending energy is a good approximation for a wide range of angles, however, it does not describe the appropriate chemical behavior of

all systems, for instance, systems where the existence of multiple equilibrium values is possible.

Dihedral torsion is characterized by their periodical behavior, with more than one minimum of energy, and therefore requires a more complex equation to describe it. There are many dihedral interactions between groups of four atoms linked by harmonic springs: the 1st and 4th atoms rotate around the bond established between the 2nd and 3rd atoms. Each time a complete rotation occurs, there's a periodic behavior best described by a Fourier expansion of a sinusoidal potential.

$$E_{dihedrals} = \sum_{dihedrals} K_{\varphi}[1 + \cos(n\varphi - \gamma)]$$

Equation 5

This equation requires the following parameters to calculate the potential energy of dihedral angles: K_{φ} represents the maximum energy of the torsion, n is the number of minima of a complete torsion, φ is the current torsion, and γ corresponds to the phase with the lowest potential energy.

Improper torsions, just like dihedral torsions, involve four linked atoms, in which one of them is bonded to the other three. This results in torsions stemming from the strain of π-molecular orbitals formed by the binding of sp^2-hybridized atoms. The potential energy associated with improper torsions is also described by a quadratic equation, in which φ is the angle formed by two non-concurrent vectors in the plan:

$$E_{\substack{improper \\ torsions}} = \sum_{\substack{improper \\ torsions}} \frac{1}{2} k_{\varphi}(\varphi - \varphi_0)^2$$

Equation 6

Non-bonded terms serve to approximately describe the behavior of electrons, which are not explicitly considered in MM methods, through electrostatic and van der Waals interactions.

Electrostatic interaction is described by Coulomb's potential and depends on the partial charges of two non-bonded atoms (q_i and q_j), vacuum permittivity (ε_0) and the distance between the two non-bonded atoms (r_{ij}).

$$E_{ee} = \sum_{nonbonded} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$

Equation 7

As for the long-range van der Waals interactions, the Lennard-Jones potential describes it considering both Pauli repulsion and London dispersion. Pauli repulsion is based on Pauli exclusion principle that states that no two electrons can occupy the same quantum state. That translates into the notion that when the distance between two atoms decreases, the repulsion between them increases, impeding further approximation. As for London dispersion, it refers to the temporary dispersive force that results when electrons in two adjacent atoms occupy a position that forms instantaneous and induced dipoles. Both Pauli repulsion and London dispersion decay very rapidly with distance, and Lennard-Jones potential describes it with a 12-6 potential between atoms i and j, at a distance r_{ij}, as follows:

$$E_{vdW} = \sum_{\substack{nonbonded \\ pairs\ ij}} \varepsilon_{ij} \left[\left(\frac{R_{min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min,ij}}{r_{ij}} \right)^{6} \right]$$

Equation 8

The term ε_{ij} describes the minimum potential energy, and $R_{min,ij}$ the distance at which the energy reaches the minimum. The positive factor with a power of 12 represents Pauli repulsion, while the negative factor with a power of 6 represents the London dispersion.

As both Coulomb's and Lennard-Jones' potential are slowly converged sums considered to infinite, a truncation factor is often used in Molecular Dynamics (expanded in the next section) to account for it.

2.3.2. Molecular Dynamics

Optimized structures often correspond to how biomolecules are found in nature. So, one of the key steps of the modeling process is structure refinement and optimization through **energy minimization** of the system.

Often, the starting point of an *in silico* study is an X-ray crystallographic structure, which requires the subsequent addition of hydrogen atoms, solvent molecules, and counter-ions to neutralize the net charge of the system. However, these additions may result in bad contacts and steric hindrance, which can be solved through the gradual relaxation of the system and energy minimization.

In the following chapters, the subject of study is enzyme-substrate systems with solvent. At the beginning of each new work, a standard procedure is to minimize the energy of the system as the following protocol:

i) minimization of solvent molecules;
ii) minimization of hydrogen atoms;
iii) minimization of protein backbone;
iv) minimization of the full system.

In the case of extensive modeling being performed in the system at study, additional steps might be introduced to the energy minimization protocol, targeting modeled regions such as loops or missing segments in the initial X-ray crystallographic structure.

With the energy minimization protocol completed, the system reached a minimum in the potential energy surface (PES). Energy minimizations modify the system's geometry, exploring the PES until an energy minimum for the system (a stationary point) is found, and perform this iteratively over a finite time period. However, at this point, the system's conformation is only one of many possible minimum conformations. Therefore, for further sampling of the system's PES, a **Molecular Dynamics (MD) simulation** is carried out. In this protocol, an initial set of coordinates for the system,

usually given by experimental data such as energy minimized X-ray crystallographic structures, will be subjected to randomly assigned initial velocities, at a chosen temperature, producing a series of time-correlated microstates that will result in a trajectory over a period of time. The MD simulation is a way to explore the ensemble of possible microstates along a period of time defined by a series of finite time steps.

The acceleration of each particle, which determines how it moves from one position to another, is calculated by:

$$\alpha = \frac{-1}{m} \frac{\delta V}{\delta r}$$

Equation 9

α, the acceleration, is calculated based on the mass m of an atom at the position r, and the V is the potential energy function from the force field equation. The new positions of the atoms, at each time step, are obtained using the acceleration α as follows:

$$x = x_0 + tv_0 + \alpha t^2$$

Equation 10

Where x is the new position, t the used time step, with v_0 initial velocities, and x_0 the initial position of the atom. The iterative application of these equations provides the trajectory of a given atom for a given timescale. As the potential energy calculation depends on all atoms, and therefore on the acceleration too, to save up computational resources, AMBER utilizes the *leap-frog* algorithm to solve these equations, updating the positions and acceleration of all atoms using the velocities of the previous step [179].

To guarantee an effective exploration of the conformational space, a series of initial parameters is key to establishing the MD simulation conditions. Firstly, it is necessary to establish a sufficiently long duration of the MD simulation and an adequate **integration time step** (which corresponds to t in Equation 10). The smaller the time step, the more accurately the MD simulation will reproduce the system's behavior

while increasing computational demands. On the opposite, the bigger the time step, the less accurate the calculations. Thus, for a given system, there needs to be a compromise between the accuracy and length of the simulation. Generally, the time step in which the forces are calculated must be at least ten times smaller than the time of the fastest molecular vibration in the system – typically hydrogens, the lightest particles. Hydrogen vibrations occur with a frequency of ~10^{-14} s^{-1}, and as the time step should be ten times smaller, it must be in the femtosecond order ~10^{-15} s^{-1}. So, considering an MD simulation of 1 nanosecond (10^{-9} s^{-1}), it will require approximately 10^{-6} time steps.

To improve the length simulation and computational cost ratio, it is beneficial to consider the bonds with hydrogen atoms. As the stretching vibration of the hydrogen atoms has little influence on the properties measured by MD simulations, the SHAKE algorithm [180] is used as an approximation to constrain all bonds involving hydrogen, and thus allow for more extended simulation periods without additional computational cost. Therefore, excluding hydrogen vibration, the second-fastest vibrations are C-C bonds, at a frequency of 20-30 fs, which can be simulated with a time step of 2 femtoseconds.

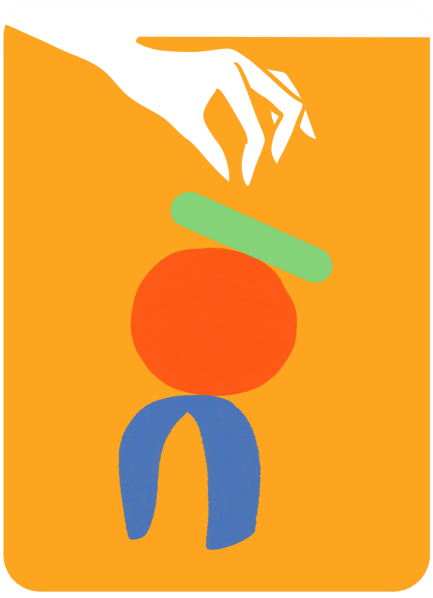


Figure 4| Balance is key for a successful MD simulation. MD simulations depend on a fine balance between the length of the simulation, the size of the system and the accuracy of the method. Illustration © Mafalda Salgueiro.

The solvent is a key player in the behavior of biomolecules. Hence, the second factor to consider is the **solvent**, which will directly influence the size of the system. In the present thesis, systems were modeled to reflect physiological conditions, and therefore the chosen solvent is water. In the context of an MD simulation, water molecules are represented by TIP3P and TIP4P, which have been parametrized to reproduce the properties of water molecules [181].

A realistic model of a solution entails a large number of solvent molecules to be included in the system along with the solute. To best mimic the behavior of bulk solvent molecules in the MD simulation, it is necessary to take into account that in the context of simulation, *i)* some solvent molecules will be at the boundary between solute and solvent, *ii)* others will be within the bulk of the solvent, and *iii)* a large number will be at the edge of the solvent and the surrounding vacuum. To circumvent this, **periodic boundary** conditions (PBC) are used so that ten to hundreds of thousand solvent molecules are needed to reproduce the physical properties of the bulk solvent. In sum, in this method, the system made up of the solute (enzyme) and solvent (water molecules) is placed into a box, for instance, of rectangular or octahedral geometry. The box is repeated *ad infinitum* to recreate bulk solvent. Considering the bulk solvent created by PBC and its infinite contribution of solvent water molecules to the system's potential energy through electrostatic and van der Walls interactions (Equations 7 and 8, respectively), there will be an increase in the computational cost of the MD simulation. To circumvent the extent of van der Walls interactions contributions, a cut-off distance is applied to impose a limit, generally 10 Å. Beyond that, a continuum damping function may be used. Regarding electrostatic interactions, best described by Coulomb's potential (Equation 7), neglecting the interacting beyond the cut-off 10 Å might originate artifacts. To circumvent this, the Particle Mesh Ewald (PME) [182] method is considered best suited for the treatment of electrostatic interactions beyond the cut-off in periodic systems.

Each phase of an MD simulation can be characterized according to state functions: the number of particles (N), volume (V), temperature (T), pressure (P), total energy (E), among others. The choice of state function will determine the **ensemble** produced throughout the MD simulation. In the present thesis, the chosen ensemble conditions

were canonical (NVT) and isobaric-isothermal (NPT). NVT ensemble stands for a number of particles, volume, and system temperature to be maintained constant. It is applied post-energy minimization and at the beginning of the MD simulation to maintain a constant volume and prevent system expansion due to unresolved repulsive forces. NVT is followed by NPT, which stands for a constant number of particles, pressure, and temperature. It is the chosen ensemble to carry out the MD simulation due to its ability to closely reproduce ambient temperature and pressure conditions.

The choice of **temperature** is also critical to the quality of the MD simulation of the modeled system, so a thermostat is employed to maintain the temperature constant throughout the simulation. To achieve the desired temperature at which the MD simulation will be performed, usually at the NVT stage, there is an incremental increase of temperature over a set period of time, over which the thermostat will be responsible for gradually adding or removing energy to or from the system with an appropriate time constant. The Langevin thermostat [183] is often applied in MD simulation, as is the case in the present thesis.

Overall, with a brief description of the most relevant aspects of MD simulations, it is easy to comprehend the advantages that MD simulations offer to model and understand the behavior of macromolecules such as enzymes. Furthermore, several software packages allow us to perform the MM calculations entailed in these MD simulations, and in the present thesis we opted for AMBER 12 package [179].
However, MM and, consequently, MD simulations have inherent limitations, especially in describing chemical reactions involving making or breaking chemical bonds. Hence, the field of QM is explained in the following section.

2.4. Quantum Mechanics

The turning of the twentieth century marked a shift in many fields, and physics was no exception. Until then, particles with mass and waves were studied in distinct disciplines: classical mechanics studied mass, and Maxwell electromagnetism studied

waves. But Schrödinger's equation revolutionized past conceptions and combined both waves and particles physics to describe the behavior of subatomic particles, such as electrons. Electrons are essential to describe chemical reactions that occur at the active site of an enzyme, as these are involved in the making and breaking of chemical bonds.

In the field of QM, electrons are treated as particles with mass and as waves – the so-called particle-wave duality. This postulate considered that when electrons are treated as waves, they will propagate as such, contradicting the classic mechanics notion of trajectory as defined by Newton's laws. In QM, the full information about a system is described by a wavefunction (Ψ), and Schrödinger's equation (Equation 11) allows for the determination of the energy for any system by solving the equation.

$$\hat{H}\Psi(x,t) = i\hbar\frac{\delta}{\delta t}\Psi(x,t)$$

Equation 11

In Equation 11, we see Schrödinger's equation, where i represents the imaginary unit, $\hbar$ corresponds to the Plank constant divided by 2π, t is time, and $\hat{H}$ is the Hamiltonian operator of the system. The wavefunction can be used to describe the position of the particles x of the system along the time t but, opposing the molecular mechanics notion of the exact positioning of particles, solving the wavefunction will provide the probability amplitude of finding a particle x in a given space and time.

However, solving Schrödinger's equation poses too big of a barrier to studying enzymes using computational methods. By considering Schrödinger's equation in its time-independent form and employing the Born-Oppenheimer approximation, the calculations become feasible for a small set of atoms. Time-independence can be applied since we are studying stationary states where energy does not correlate with time, resulting in the simplified time-independent Schrödinger's equation where x corresponds to the coordinates of the all atoms in the system:

$$\widehat{H}\Psi(x) = E\Psi(x)$$

Equation 12

In equation 12, $\widehat{H}$ stands for the Hamiltonian operator that operates on Ψ, a mathematical wavefunction, to yield the total energy E of the system multiplied by that same wavefunction Ψ.

The next crucial approximation is the Born-Oppenheimer approximation which treats electronic motion independently of nuclei motions, as the latter move much slower due to their higher mass than the former. As such, the movement of the electrons and nuclei can be decoupled, the kinetic energy of nuclei can be ignored, the nucleus-nucleus repulsion can be considered constant for a given geometry, and the electronic energies calculated for fixed nuclear positions. This results in treating the nuclei as stationary, and the nucleus-nucleus repulsion can be treated as a parameter and added *a posteriori*:

$$\widehat{H} = \sum_{i=1}^{n} \frac{-1}{2}\nabla_i^2 + \sum_{i=1}^{n}\sum_{K=1}^{N} \frac{-Z_K}{|r_i - R_K|} + \sum_{i=1}^{n}\sum_{i>j}^{n} \frac{1}{|r_i - r_j|}$$

Equation 13

The first term is the sum over the kinetic energy of the electrons of the system, the second term the electrostatic interactions between electrons and nuclei, and the third term the sum of the electrostatic interactions between pairs of electrons. In the second term, the Z_K stands for the atomic number. Finally, as aforementioned, the constants concerning the repulsion between stationary nuclei are added *a posteriori*.

The Born-Oppenheimer approximation led to the concept of potential energy surface (PES). The PES corresponds to a surface defined by the electronic energy over all possible nuclear coordinates. At this stage, the third term, which refers to the sum of the electrostatic interactions between pairs of electrons, can only be solved for a model with one electron. As QM must be able to treat systems with several atoms, more approximations would be necessary.

Several methods have been developed to numerically solve Schrödinger's equation, including the Hartree-Fock (HF) method, and following post-HF methods that offer accurate numerical solutions at the expense of increasing computational power. However, in the present thesis, we opted for Density Functional Theory (DFT), another alternative to approach bigger systems like enzymes at a reasonable computational cost.

2.4.1. Density Functional Theory (DFT)

In a nutshell, DFT methods describe molecules' geometry and energy, resorting to electron density instead of wavefunctions. The Thomas-Fermi model, which was the predecessor of the modern DFT theory, defends that the electronic energy of the system can be obtained taking into account the total electron density at each point in space, replacing the wavefunction altogether. Side by side, the stark difference between the idea behind the DFT method vs. wavefunction is the number of variables each requires: the wavefunction depends on four coordinates per electron (three spatial variables (x, y, z) for each electron, plus its spin variable), while the DFT only depends on the three spatial coordinates of the electron. This results in reduced computational time.

DFT has finally proposed through the work of Hohenberg-Kohn, and their two theorems which state that *i)* there is an electron density functional that when applied to non-degenerated systems can provide the exact ground state energy, and *ii)* that the electron density that gives the lower energy is the ground-state exact density. However, the Hohenberg-Koh theorems exhibited a few shortcomings, particularly that the exact form of the density functional is unknown. Therefore, the Khon-Sham formalism proposed a method to solve this problem by finding the lowest energy electron density by splitting the density functional of the system into two parts: an exact part, common to all density functionals, that considers electrons as non-interacting particles, and a second part, named exchange-correlation term, which corrects the previous part by accounting for the electron-electron interaction.

$$E_{tot}[\rho] = T_s[\rho] + E_{ne}[\rho] + J[\rho] + E_{XC}[\rho]$$

Equation 14

The first three terms of equation 14 can be solved explicitly, while the fourth term – the exchange-correlation functional $E_{XC}[\rho]$ – is unknown and must be approximated to predict the system's total energy.

Thus, the electronic ground-state energy of the system can be written as:

$$E_{tot}[\rho] = -\frac{h^2}{2m}\sum_i \int \Psi_i^* \nabla^2 \Psi_i d^3r + \int V(r)n(r)d^3r + \frac{e^2}{2}\iint \frac{n(r)n(r')}{|r-r'|} d^3r d^3r' + E_{XC}[\rho]$$

Equation 15

The density functional on Equation 14, with ρ as a variable for the electronic density, can be divided into four terms as seen in Equation 15: the first term $T_s[\rho]$ describing the kinetic energy of electrons, the second term $E_{ne}[\rho]$ referring to the Coulomb interactions between the electron density and the nuclei, the third term $J[\rho]$ accounting for the repulsion between electrons, and lastly, the fourth term $E_{XC}[\rho]$ is the exchange-correlation term that accounts for the correlation effects and corrections to the kinetic energy and Coulombic repulsion terms.

The accuracy of the DFT method depends on how accurately the exchange-correlation functional can be determined. Several approximations have been developed to solve the exchange-correlation functional. The biggest challenge with DFT methods is exactly with which functional will best describe the exchange-correlation functional $E_{XC}[\rho]$.

2.4.2. Exchange-correlation density functionals

The simplest approximation available to describe the exchange-correlation functional is based only on the electron density, called Local Density Approximation (LDA). In LDA, the energy depends only on the density in a homogeneous electron gas at each point in space. In other words, the electron feels the electron density produced by the

remaining electrons in every point in space in the system as if the density was the same in each part of the system. However, this approximation fails to describe the electronic density changes in, for instance, biomolecular systems.
Based on LDA, other approximations came to light, one of the first being the Generalized Gradient Approximation (GGA) that incorporates electron density while also considering its gradient. This approximation is more apt to describe inhomogeneous molecular densities, which have been improved with hybrid functionals. Hybrid functionals combine GGA functionals with HF exchange. The predominantly used hybrid functional is B3LYP [184–186], which has been employed to study the enzymatic catalytic mechanism of GE, as described in the present thesis.

2.4.3. Basis sets

In a QM calculation, basis sets are groups of functions that are linearly combined to represent molecular orbitals. The larger the basis set, the more accurate description of the molecular orbitals. However, one cannot use an infinite number of functions. Instead, just as is the case of the ethos of QM/MM methods, there needs to be a compromise between accuracy and computational power expenditure. Additionally, the basis set must be relatively transferrable - the basis for a given atom is flexible enough to be used for that atom in various bonding environments where the atom's hybridization and local polarity may vary. Considering this, there are two main types of basis functions: Slater type orbitals (STO) and Gaussian type orbitals (GTO).

A wavefunction for an isolated atom or molecule will exponentially decay to zero over great distances between the nucleus and the electron ($r \rightarrow \infty$). Its maximum energy will be at the nearest distance to the atom nucleus ($r = 0$). STO basis function best describes the electron density around an atom. They are exponential functions, and their exponential dependence on the distance between the nucleus and electron (r) represents the exact orbital for a hydrogen atom. However, despite being accurate in their description of mono and diatomic systems, their application in bigger systems is set back by the demand for computational power.

An alternative to STO basis is GTO, which greatly differs from STO at $r = 0$ and ($r \rightarrow \infty$) because the wave function in GTO is squared, whereas the STO wave function is not. Yet, despite GTO difficulties in representing the behavior of electrons near the nucleus, it offers the advantage of being less demanding in terms of computation. To improve the GTO description of the atomic orbital, there is the contracted GTO (CGTO). CGTO is a linear combination of primitive Gaussian functions that offer more accuracy and whose behavior approximates to STO. The minimal basis set consists of a CGTO that describes both core and valence electrons of the system, but this is a rough approximation because they fail to describe the valence orbitals accurately. So, in CGTO, a linear combination of primitive Gaussian functions can be used to improve the description of electrons and their behavior according to the orbitals they occupy. The earliest CGTO basis sets were constructed from *n*-GTOs that best fit the desired STO.

Just as electrons are subdivided into core electrons and valence electrons, that same principle can be applied to the basis set. The core orbitals will describe inner electrons, and valence orbitals will best describe the outer electrons. Pople Basis Set is *n-abc*G split valence type basis set. *n* indicates how many GTO are used to describe the core orbitals of the system. Then, *a, b,* and *c* indicate the number of functions that the valence orbitals are split into, and how many are used to describe it. If it is a *n-ab*G, then the basis set is called double-zeta. If it is a *n-abc*G, then the basis set is called triple-zeta because it represents a triple split orbital. Let's consider 6-31G, the double-zeta basis set used in our calculations. In **6-31G**, the **6** represents the 6 primitives that were combined into a CGTO to describe the inner shell (1s orbital), and the valence shell is split into an inner and outer part (double zeta) – the inner using **3** primitive Gaussian functions and the outer using **1** primitive Gaussian function.

To further improve the basis sets, two types of functions are added: polarization functions and diffuse functions. Taking the 6-31G double-zeta basis set as an example, the inclusion of polarization functions results in 6-31G**** or 6-31G(**d,p**) – the nomenclature being added after the G. Polarization functions improve the flexibility of the orbitals. The first * or d refers to added polarization functions for heavy atoms,

while the second * or p refers to polarization functions added to the hydrogen atoms. As for the diffuse function, they are added to improve the description of the orbitals when it is expected that the electron density slowly decays far from the nucleus. Their addition is indicated by **+** or **++**. The first + indicates that diffuse functions were added for heavy atoms, while the second + indicates the addition of a *s* diffuse function to hydrogen atoms. The addition of diffuse functions is beneficial when describing systems with anions. Highly excited electronic states and large atoms tend to have more diffuse electronic density and even a weakly bound electron sparse out from the nucleus.

2.5. Hybrid QM/MM Methods

Until the late twentieth century, only small systems were eligible to be studied with QM methodologies. This curbed back the computational study surrounding enzymatic catalysis and other complex molecular systems. The combination of quantum mechanics and molecular mechanics was first developed by Karplus and Warshel in the 1970s, resulting in the seminal paper *Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme* [187]. In 2013, alongside Michael Levitt, Karplus and Warshel were awarded the 2013 Nobel Prize in Chemistry for the *Development of Multiscale Models for Complex Chemical Systems* [188].

At the core of QM/MM methodologies lies the combination of quantum mechanics and molecular mechanics, each applied to different segments (layers) of the simulated system. The most accurate and computationally demanding QM methods are used on the active site residues of the enzyme because they can describe electronic rearrangements. On the other hand, MM methods treat the remainder region outside of the active site, contemplating the surrounding and environmental effects. Altogether, the hybrid QM/MM approach allows for an atomistic description of the enzyme-catalyzed chemical reactions by characterizing its transition states and intermediary species.

For QM/MM methodologies, the total energy of the *real* system is defined as $E_{QM/MM}$, and it is represented by the sum of the energy of the *model* system treated with QM methods E_{QM} (active site region, also called *QM layer*), plus the energy of the *environment* treated with MM methods E_{MM} (also called *MM layer*), and plus the interaction between the QM *model* system and the MM *environment* system E_{QM-MM}:

$$E_{QM/MM,real} = E_{QM,model} + E_{MM,environment} + E_{QM-MM}$$

Equation 16

This description of QM/MM methods is called the additive scheme, where the E_{QM-MM} energy of interaction between the *model* and *environment* is particularly tough to determine because it includes: *i)* bonded interactions for the covalent bonds located at the boundary between the QM and MM treated regions; *ii)* non-bonded interactions (electrostatic and van der Walls interactions).

An alternative approach to QM/MM methods is the subtractive scheme, developed by Vreven, Morokuma et al. in the ONIOM method (Our Own N-layered Integrated Molecular Orbital and Molecular mechanics) [145,189,190]. In the ONIOM method, the system can be divided into *N* layers (usually two, but sometimes more) that can be treated with equal or different methods (QM/QM or QM/MM). The many combinations possible of *N*-layers and applied methods make ONIOM a highly attractive, tailored method. We have chosen ONIOM to carry out the enzymatic catalysis studies shown in this thesis.

2.5.1. Subtractive ONIOM scheme

$$E_{ONIOM} = E_{QM,model} + E_{MM,real} - E_{MM,model}$$

Equation 17

In the ONIOM method, energies are calculated through a subtractive, or extrapolative method. The total energy of the system E_{ONIOM} is the result of the QM energy of the *model* system $E_{QM,model}$, that encompasses the reactive active site region, plus the

MM energy of the total *real* system $E_{MM,real}$ minus the MM energy of the *model* system $E_{MM,model}$ – thus being called the subtractive method. By subtracting the MM energy of the *model* system (active site region) $E_{MM,model}$, we can calculate the energetic contribution of the *environment* using a less demanding method. However, the calculation of the energetic contribution at the boundary between the QM region and MM region was yet to be improved.

In the original subtractive scheme, the interaction at the boundary region is treated with the MM forcefield, thus being called the mechanical embedding scheme (ME-ONIOM). But this approach does not contemplate the mutual polarization between the QM and MM regions; the electrostatic embedding scheme is used to solve this.

2.5.2. Mechanical embedding, electrostatic embedding and link atoms

The **mechanical embedding (ME)** approaches the system as if the QM layer exists in a vacuum and therefore it is not polarized by the environment region. The interactions at the boundary region between the QM and MM layers are calculated at the MM level, so the MM level is influenced by the QM layer. The main limitation in mechanical embedding is that it describes QM-MM the interactions at the boundary region as classical Coulomb interactions between point charges: atomic charges are assigned to the *model,* and *real* systems, usually defined with the force field (in our case, AMBER) and these are not updated despite geometry changes throughout optimization steps. This impacts the calculation quality because the electronic state and charge distribution can change during geometry optimization, mainly if large charge transfers (such as bond breaking or making) occur. Nonetheless, it is possible to recalculate the charge of the high layer during the study of the reaction, at each geometry optimization step, using the electrostatic potential (ESP) or restrained electrostatic potential (RESP) [191].

Interestingly, in the Glucuronoyl Esterase enzyme work, our first approach was ME-ONIOM. The GE active site is located near the protein's surface, in contact with water molecules from the solvent, and therefore in a polar environment. The ME-ONIOM

approach failed to describe the QM region's charge density polarization by the MM region's electrostatic environment, resulting in unsuccessful calculations.

Despite the alternative to recalculating the high layer charges using ESP or RESP at each geometry optimization step, we opted for the electrostatic embedding scheme, which granted more reliable results.

The **electrostatic embedding (EE)** scheme is similar to the additive scheme because it uses a coupling scheme for the Coulomb interactions. In broad terms, in the EE scheme, the high layer is subjected to the electrostatic field generated by the MM point charges, thus being polarized, and it is included in the high layer Hamiltonian (QM level) as an external potential, hence the similarity with the additive scheme.
Alongside electrostatic interactions, van der Walls interactions are also important non-bonded interactions. These are treated similarly in the EE-ONIOM and ME-ONIOM schemes.

The main advantage of the EE-ONIOM scheme is that it allows for the electronic structure of the *model* system (high layer) to adapt to the surrounding environment (low layer), and therefore the electronic density of the high layer is much closer to reality than when calculated using ME-ONIOM. However, the higher accuracy of the EE-ONIOM method requires increased computational time, so it is important to balance the size of the QM layer and the time required for the calculation.

Another vital aspect of QM/MM methods is the treatment of the boundary between QM and MM layers. When preparing the system to be studied with QM/MM ONIOM methods, it is necessary to define which residues, and therefore atoms, will be studied at the QM level (high layer), and during this process, it will be required to assess which covalent bonds will need to be cut in between the two layers. Once the decision is made, always considering the compromise between the size of the QM layer and its computational expense, the user has to consider how to best deal with the now dangling bonds at the boundary between QM and MM layers – this has been a critical issue that undermines the accuracy and applicability of the combined QM/MM method.

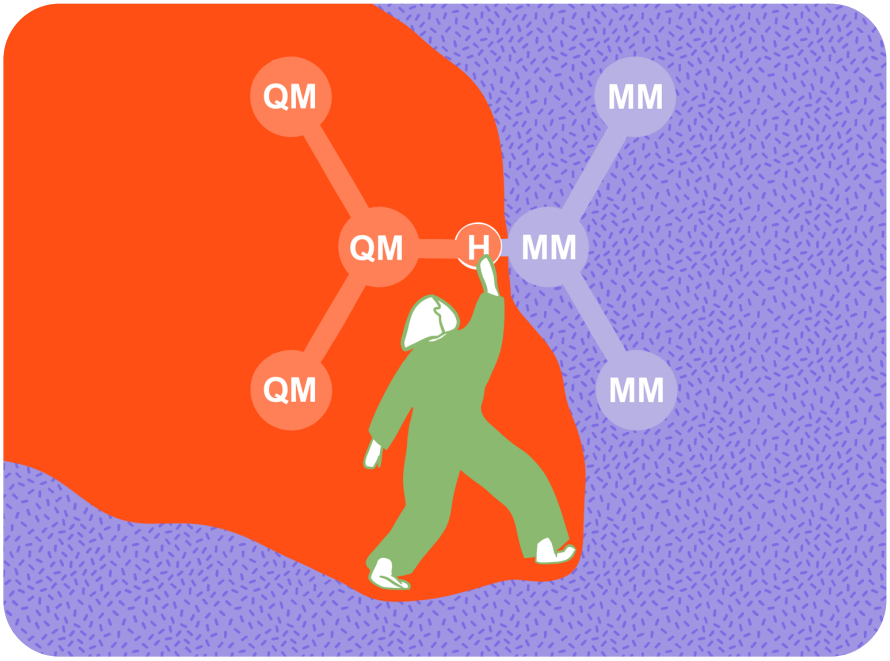


Figure 5| Link atoms. Link atoms are placed between the atom of the low layer and the atom of the high layer, and covalently bound to the last one (high layer). The goal of the link atom is to cap the free valence of the active part in the high layer. Illustration © Mafalda Salgueiro.

One of the many [192] approaches to describe the **QM/MM boundary** across the covalent bonds, and the one used in the present thesis, is **link atoms**. Link atoms are placed between the atom of the low layer and the atom of the high layer and covalently bound to the last one (high layer). The purpose is to introduce an additional atomic center, usually a hydrogen atom, to cap the free valence of the active part in the high layer. This eliminates the existence of a free radical in the high layer. As for the MM layer, the link atom will play its role in keeping the MM bond, angle, dihedral terms to maintain the system's geometry.

2.5.3. QM/MM geometry optimization

Having understood the theory behind the MM and QM method and some of the aspects of the hybrid QM/MM method, it is essential to understand how the hybrid QM/MM works.

As previously mentioned, through the exploration of the PES, it is possible to find the location of stationary points: transition states and minima, which include reagents, intermediates, and products. At each exploration of the PES, it is necessary to modify the coordinates of thousands of atoms, with many degrees of freedom, optimizing energies to find a new minimum. When treating large systems such as enzymes, hybrid methods that partition the system into different layers allow for the study with great accuracy and without excessive computational expense. In the ONIOM method, the exploration is carried out using the microiterative optimization strategy: the QM region and the MM region are optimized alternately, allowing for the optimization of minima and transition states, even when using EE. So, the process is: the geometry of the MM layer is fully optimized for a given conformation of the QM model. Once the MM model achieves zero energy, the QM can move one step forward in what is called macroiteration. During the macroiteration, there are as many steps as required (under a limit of steps defined in the input file) until the system achieves a stationary point on the PES. In the software used for this thesis, Gaussian 09, the terminology employed is whether or not the optimization has converged to a value under a given threshold (also defined in the input file).

2.5.4. Best practices for the use of hybrid QM/MM methodology

If no modeling is required, one can often use an X-ray structure, followed by an energy minimization protocol. However, when extensive modeling is necessary, it is crucial to first sample as many possible conformations throughout an MD simulation. Given the vast array of possible paths over the PES, it is vital to acknowledge that exploring a single reaction path is a limited, counterproductive approach. Biomolecules are not in a static state. Through extensive optimization, sampling, and further optimization of the initial structure, it is possible to identify conformations that are reliable starting points for the QM/MM study. Another factor that plays into the quality and accuracy of the results obtained with QM/MM is the choice of residues and atoms for each layer – high and low. Thus, before committing to a single reaction path, it is advised to explore as many conformations and reaction coordinates as possible while being mindful of the available literature on the subject of the study.

2.5.5. How to study the catalytic mechanism of an enzyme using QM/MM?

A study of the catalytic mechanism of an enzyme entails many steps, each with its specific goal. Some of the critical aspects of QM/MM, particularly ONIOM methodologies, were briefly introduced in the previous sections, giving an overview of the theoretical foundations for this approach. As an attempt to create a protocol for future students to rely on, here follows a streamlined and standardized approach for a QM/MM study, which was loosely applied to the study of the catalytic mechanism of GE (Chapter 4). Thus, as cautionary advice for future applications of the following protocol, one should always conduct scientific research with a critical and curious spirit while always considering the specifics of the system subject of the study.

Thus, before initiating the study of an enzyme, it is key to understand its role in the broader context and perform extensive literature research and reading. Scrutinizing available literature on a given topic will help identify knowledge gaps and identify questions yet to be answered. In this process, the scope of the forthcoming study will be defined, and a starting point will be clear.

Preparation of the system

While conducting the literature research to begin a QM/MM study, it is key to look for X-ray crystallographic structures available at the PDB database [43,136] for the target enzyme. The X-ray structure will provide the initial coordinates for modeling the enzyme reaction, so the user must always analyze the available structures. Factors to take into account are: *i)* good resolution of the X-ray structure, preferably between 1.0 and 2.5 Å. Above 2.5 Å, the backbone and bulky sidechains are still mostly clear, yet systematic errors, such as missing or misplaced atoms, may arise, alongside increasing coordinate error; *ii)* comparing the enzyme's its primary sequence (retrieved from UniProt [193]) with the sequence of the PDB file, checking for missing fragments that failed to be crystallized in the process (this information is available at the header of the PDB file), and checking for mutated residues (a standard procedure enhancement of crystallization [194,195]); *iii)* presence of ligands or substrate analogs.

Once the X-ray structure has been selected to initiate the study of the catalytic mechanism of the enzyme, and after close inspection of the PDB file, the user can edit it to remove crystallization buffer molecules and to correct mutations. Hydrogen atoms are also usually missing from the X-ray structure and need to be added *a posteriori*. To add the missing hydrogen atoms, there are different tools available to predict local pK_A values of each residue, such as PROPKA [196] or the ProteinPrepare tool from the PlayMolecule online server [197]. However, using these tools still requires the commitment of the user to visually inspect the structure, taking into account surrounding residues for each titratable residue, because the wrong protonation can affect the results of the catalytic mechanism study.

Often, the PDB structure does not contain the desired ligand (for instance, the substrate molecule) mainly because the enzyme turnover rate is so rapid that it is challenging to crystallize it with its natural substrate. So, it is often the user's task to model the ligand accordingly. This can be accomplished either by *i)* superimposition and/or modification of the PDB structure's ligand to assembly the desired ligand while maintaining its position at the active site, or *ii)* through modeling of substrate molecule followed by docking of the substrate into the active site. Substrates' parameters are not available in common forcefields, such as AMBER's [198] FF99SB-ILDN [199] (developed for protein and nucleic acid systems) or GAFF [200] (General AMBER Force Field, developed to describe a broader range of molecules than available protein and nucleic acid force fields). So, in either case, substrates need to be parametrized – charges, vdW parameters, and parameters for intramolecular interaction – for the assembly of the enzyme-substrate system.

Once the enzyme-substrate system is completed, with the ligand in the reaction-prone pose (based on previous data), and all the mutations and missing residues corrected, it is time to add counter-ions to solvate the system, followed by the energy minimization procedure. As previously explained, the energy minimization step is crucial to system refinement and optimization. It will prevent clashes and bad contacts in the system that originated in the modeling phase, and it will allow the gradual relaxation of the system and energy minimization.

As the energy minimization of the system does not offer the possibility of exploring the many degrees of freedom of the enzyme, starting the QM/MM study at this stage will result in a narrowed sampling of the enzyme's conformational space. However, there are many options to generate an extended conformational exploration – for instance, using MD simulations to capture several snapshots of the enzyme-substrate complex with different geometries [201,202]. Another route to guarantee a broad exploration of the conformational space is performing extensive MD simulations with multiple replicas. Furthermore, this can be combined with clustering to define the most prevalent poses of the enzyme-substrate complex. In chapter 3, we describe the use of extensive modeling and ample MD simulations to ensure a reliable enzyme-substrate complex as the starting point of the QM/MM study.

Establishing QM and MM layers

With a chosen enzyme-substrate pose, it is possible to divide the system into separate layers, studied at different levels of theory – QM methods in the high layer, MM methods in the low layer. As explained, QM calculations have a higher demand of computational power, and therefore one must choose wisely the residues to be studied with QM.

Firstly, one must select the active site residue directly involved in the catalyzed reaction and the substrate. Counting the number of atoms is needed at each selection because one must keep it at a reasonable number. Using DFT to study enzymatic catalysis allows for a QM layer with over one hundred atoms, above which calculations might become too computationally demanding. For instance, if the substrate is a molecule with many atoms, one could assume that a portion of it might be representative for the catalysis, with the remainder in the MM layer.

Once selected the catalytically active residues and substrate, it is crucial to note which intramolecular interactions are established around the active site – hydrogen bonds, ionic bonds, or even disulfide bonds as seen in Chapter 5, where we included a disulfide bond in the QM layer due to its role of keeping the integrity of the active site.

These may also be added to the QM layer, but a compromise must always be made. Hence, a plausible solution is to only include parts of the residues, not both the side chain and backbone of the residue, which brings us to the next question: where to cut the residues? The user may apply their own criteria, but there are two particular cases to be mindful of:

1. Boundaries between layers may never be placed within aromatic rings, or delocalized double bonds
2. Atoms linked by double and triple bonds must be kept in the same ONIOM layer (either QM or MM, never separately).
3. Atoms that undergo geometry or electronic adjustments during the reactions must be considered for the QM layer.

Once decisions have been made regarding the QM layer, it's essential to decide how to treat the environment. The MM layer must aptly represent the active site surroundings. The best way to ensure it is to keep the whole protein plus a hydration shell in the MM layer. This will ensure the interactions between the solute (enzyme) and the solvent are maintained and accounted for.
Finally, before starting the QM/MM geometry optimizations, the next step is to define the radius after which atoms will be restrained – commonly called "freezing." Restraining a portion of the enzyme during QM/MM calculations prevents problems arising during geometry optimization due to too many atoms.

Choosing the level of theory

As previously explained, DFT methods offer a decreased computational time due to the reduced number of variables (solely three spatial coordinates for the electron) in comparison to wavefunction-based methods (which contemplate the additional spin variable). Hence, DFT methods are currently the preferred level of theory for QM calculations.

B3LYP has been the preferred functional to carry out geometry optimizations and explore the potential energy of reaction coordinates [203–206]. However, given the

accuracy required in single-point energy calculations, B3LYP does not perform as well [207] and other functionals are used to characterize key stationary points. Therefore, depending on the type of systems being studied, it's advisable to look for benchmarking studies available in the literature. These comprehensive studies explore a broad scope of reaction properties with different functionals and determine which ones perform best. Regarding MM calculations, we have used the AMBER force field ff99SB in our studies. Nevertheless, we encourage the user to check current alternatives such as ff14SB [208] or ff99SB-ILDN [199], both improvements from ff99SB.

How to explore the reaction coordinate?

With the enzyme-substrate complex ready for QM/MM calculations, the first step is geometry optimization. The intent is to minimize the system's energy through a series of consecutive iterative calculations until the convergence threshold is met. This first geometry optimization might take longer than usual, and therefore it is usually best to perform the optimization with ME instead of EE.

Once the enzyme-substrate complex is optimized, it is possible to explore a likely reaction coordinate for the proposed chemical reaction, either from previous experimental or computational studies, or based on canonical enzymatic mechanism. The exploration is achieved through linear scans along a reaction coordinate defined by bond length, although it is also possible to define it with angles and dihedrals. These linear scans will result in the exploration of the PES, and therefore it is possible and advisable to explore different reaction coordinates. When the exploration of a reaction coordinate through a linear scan is successful, the resulting energetic profile will display a minimum and a maximum energy value. The rule of thumb is avoiding energy maxima above 25 kcal·mol^{-1}. Despite further calculations being necessary to determine the energetic reaction barriers accurately, one must not forget that around $\Delta G‡$ 20 kcal·mol^{-1} reactions can take from seconds to several hours, while reactions with $\Delta G‡$ > 25 kcal·mol^{-1} may take days or weeks to complete.

With a clear maximum, usually refined by consequent linear scans with a smaller interval, it is advisable to optimize the transition state corresponding to the maximum

of energy. Then, one can proceed with an IRC (intrinsic reaction coordinate) calculation from the optimized transition state to obtain minima for the reactant and product or intermediate. The IRC corresponds to the minimum energy reaction pathway in mass-weighted cartesian coordinates between the transition state of a reaction and its reactant and product states. Thus, through the IRC, it is possible to find minima that do not correspond to the initial reactant and/or product.

Establishing the final reaction profile

Once the reaction path and its stationary points have been determined and optimized, preferably from IRC calculations, a frequency calculation for the optimized geometries follows. Frequency calculations provide the vibrational frequency for each vibrational mode of a molecule. Depending on the type of stationary point, frequencies will vary: for a transition state, the vibrational frequencies will be imaginary; if it is a minimum, all vibrational frequencies will be positive values.

Finally, it is advisable to employ a functional with increased accuracy and a more complete basis set for every single point energy calculation. In subchapter 4.2, we have characterized ten stationary points using high-level domain-based local pair natural orbital coupled cluster theory with single-, double-, and perturbative triple excitations (DLPNO-CCSD(T)) method [209], which performs better than DFT functional B3LYP [210] used for geometry optimization and linear scans.

CHAPTER 3. Fatty Acid Synthase

3.1. Fatty Acid Synthase: an introduction

Fatty acids are ubiquitous molecules found in all kingdoms of life and conserved across species. Fatty acids amass many roles in animals, but three of its most preeminent ones are their involvement in membrane architecture [211] and energy storage [212] – they act as building blocks in cellular membranes, and are used in the production of storage lipids such as triacylglycerols [213] – and their derivatives have signaling properties [214,215]. Hence, the fatty acid biosynthesis pathway is of extreme importance to the well-function of most living cells.

Fat biosynthesis is a process known since the nineteenth century [216]. Still, it was only in the 1950s and 1960s that a broader understanding of this process was established - fatty acid *de novo* biosynthesis occurs in the cytosolic fraction of the cell, and that the fatty acid synthesis occurs as described by Wakil and co-workers [217]:

$$\text{acetyl-CoA} + 7\text{malonyl-CoA} + 14\text{NADP}^+ + 14\ \text{H}^+ \rightleftharpoons$$
$$\text{palmitic acid} + 7\text{CO}_2 + 8\text{CoA} + 14\text{NADP} + 6\text{H}_2\text{O}$$

Equation 18

Where acetyl-CoA is the primer of the condensation reaction providing the carbons 15 and 16 of the resulting palmitate molecule, while malonyl-CoA is responsible for the remaining 14 carbons.

Regarding the molecular details of the fatty acid biosynthesis process, research performed by the Vagelos group on *E. coli* had shown that acyl intermediates were bound to the 4'-phosphopantetheine (PNS) group of ACP [218,219], the substrate-shuttling domain of FAS. Identification of the remaining catalytic domains of the FAS system followed in the succeeding decades.

The FAS systems are classified into two distinct types: Type I and Type II. Type I FAS is characterized by individual catalytic units that have merged into a single megaenzyme, expressed from a single gene. It is found in higher eukaryotes, fungi and *Corynebacterium*, *Mycobacterium*, and *Nocardia* (CMN) group bacteria [220]. Type II FAS systems' intervening catalytic units exist as independent and free-standing, creating a metabolic pathway for fatty acid biosynthesis [59].

Focusing on the Type I assembly, fungi and animal FAS systems differ in terms of organization, the number of catalytic domains, and organizational scaffolding. Type I fungal FAS is a barrel-shaped closed system, with 50% of scaffolding residues responsible for maintaining the rigidity of the megaenzyme and positioning the functional domains within the system. In the fungal Type I FAS system, the ACP domain responsible for transporting the substrate and intermediates from domain to domain is kept inside the reaction chamber, and the active sites of the megaenzyme are oriented towards the inside of the barrel. The metabolic crowding, with its functional domains packed into a 2.6 kDa structure, makes fungal FAS a highly efficient apparatus. Oppositely, animal type I FAS resembles a humanoid shape with its catalytic domains be exposed to the surrounding environment, with its ACP domain not confined to the reaction chamber, and the ACP and the β-ketoacyl reductase (KR) domains connected through a linker, and the TE domain connected to the ACP by an approximately 20 residues long linker. The behavior of human FAS was the subject of the study in subchapter 3.2. Through visual inspection of a classical molecular dynamics simulation of the complete human FAS protein, it was possible to observe the tremendous mobility offered by the linker [55]. The understanding of the overall mechanism of the Type I FAS system, as well as its substrate transport mechanism, have been two of the many questions surrounding FAS complex architecture.

From a disease-oriented point of view, the interest in FAS systems as therapeutic targets started in 1950 when the first connection between cancer development and levels of FAS was made. Since then, it was observed that tumor fatty acid synthesis depends almost exclusively on *de novo* biosynthesis by FAS, leading to the rapport of FAS as a cancer marker in patients. FAS levels have also been associated with the

development of diabetes and obesity. Understanding the physiological role, alongside catalytic mechanisms and structure of FAS systems, is key to developing efficient *in vivo* inhibition of FAS for therapeutic purposes. Furthermore, as proven by the extensive review developed and published by our group in late 2020 and available in subchapter 3.4.1. and Appendix section, the FAS enzyme has a wide array of applications in different fields, ranging from cancer therapeutics to biofuels. Thus, FAS is an undeniably promising subject for scientific research.

The PKS systems, which will be the subject of the review in subchapter 3.4.2., can produce various carbon skeletons – polyketides – from simple organic precursors. PKS enzymes, similarly to FAS systems, also exist as single megaenzymes or as individual enzymes that form a metabolic pathway [221,222].
Previous studies exploring the effects of insertion and deletion of domains into the animal FAS have demonstrated its ability to accommodate these alterations without affecting the integrity of the remaining domains. Further investigation led to Keatinge-Clay and Stroud's work in which swapped domains between type I FAS and PKS to produce a hybrid synthase [223], which is sustained by the evidence of common ancestry between PKS and animal FAS systems [224]. With the closeness between these two enzymatic systems and their potential for the synthesis of complex molecules, a review on the PKS systems was written [225] and follows on subchapter 3.4.2.

Altogether, both reviews on the state-of-the-art of type I FAS and PKS systems, alongside the hFAS homology model and modeling of hFAS TE domain, contribute to a new wave of research on these systems. However, many questions remain unanswered, such as the reasoning behind the movement of ACP along the multi-enzyme cycle: is it random or driven? Which factors determine its path and sequential selectivity? We hope we have contributed with our seminal work of homology modeling and ACP docking assays. Still, these questions will require further extensive description of the dynamics of animal FAS structure.

3.2. Modeling of Human Fatty Acid Synthase and *in Silico* Docking of Acyl Carrier Protein Domain and Its Partner Catalytic Domains

Matilde F. Viegas[a], Rui P. P. Neves[a], Maria J. Ramos[a], and Pedro A. Fernandes[a]

[a] LAQV@REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

In 2008, the three-dimensional crystal structure of mammalian FAS was solved for the first time at 3.2 Å resolution (PDB code: 2VZ8/ 2VZ9) [48]. However, this crystallographic structure was not complete – out of the seven catalytic domains found in each FAS monomer, only five were solved. Furthermore, the remaining two domains (ACP and TE) were absent in resolved the structure due to their inherent mobility – both are connected to the central scaffold of FAS megasynthase through connecting linkers, causing additional impediments to finding stalled positions when tracing electron density.

In 2018, as the result of the work presented in this chapter, a homology model of the human type I FAS was published, packing, for the first time, the 14 catalytic domains in its two monomers [55]. Moreover, in the present chapter, we share the results of protein-protein docking assays, and MD simulations carried out intending to understand the behavior of the FAS ACP domain and propose possible poses of ACP throughout the cycle of fatty acid biosynthesis, thus mapping obligatory checkpoints in its trajectory. Altogether, this published research paper laid the foundation for the ensuing works into the catalytic mechanisms of FAS catalytically active domains involved in the *de novo* biosynthesis of fatty acid molecules [226–230].

Regarding the contributions to this work, Matilde F. Viegas performed the calculations with the guidance of Rui P. P. Neves, and under the supervision of Pedro A. Fernandes

and Maria J. Ramos. Matilde F. Viegas wrote the first draft of the manuscript, and all authors have revised it through their contributions.

Viegas, M. F., Neves, R. P. P., Ramos, M. J., & Fernandes, P. A. (2018). ***Modeling of Human Fatty Acid Synthase and in Silico Docking of Acyl Carrier Protein Domain and Its Partner Catalytic Domains.*** In The Journal of Physical Chemistry B 122 (1), 77–85. doi:10.1021/acs.jpcb.7b09645

3.2.1. Abstract

Human fatty acid synthase (hFAS) is a megasynthase whose main function is de novo biosynthesis of saturated fatty acids. Interest has been drawn to this enzyme beyond its physiological role due to the association between high levels of hFAS and clinical conditions such as obesity, diabetes, and cancer. Thus, it has become an undeniably attractive pharmacological target. Until now, no crystal structure of the complete hFAS is available, hindering attempts to fully understand this protein. Using homology modeling, we built a model of the entire megasynthase, encompassing all of its domains, including the acyl carrier protein (ACP) and thioesterase (TE) mobile domains absent in the crystal structure of mammalian fatty acid synthase (FAS). On a second stage, we used data-driven protein-protein docking between the substrate shuttling domain ACP and every catalytic domain in the protein. We also propose sets of amino acids at the interface of each domain that we believe are important to favor the interaction between ACP and each domain of hFAS. After inspection, we validated each complex between ACP and MAT/KS/KR/DH/ER domains through classical molecular dynamics simulations and RMSd analysis. Additionally, we mapped the interactions between the residues at the active site of each catalytic domain and its intermediaries. In every docking, we ensured that the distance between catalytic residues and the intermediaries was maintained. Until now, there was not a complete 3D model of this megasynthase. This study is the first to present a homology model for the whole hFAS, including its two mobile domains and possible poses of ACP throughout the cycle of fatty acid biosynthesis, thus mapping obligatory checkpoints in its trajectory. Hence, we believe that these structural insights will allow for future studies of the catalytic mechanism of the overall hFAS.

Keywords: Human Fatty Acid Synthase, Homology Modeling, Protein-Protein Dockinig, HADDOCK, Acyl Carrier Protein.

THE JOURNAL OF PHYSICAL CHEMISTRY B
Article
Cite This: J. Phys. Chem. B 2018, 122, 77−85
pubs.acs.org/JPCB

Modeling of Human Fatty Acid Synthase and *in Silico* Docking of Acyl Carrier Protein Domain and Its Partner Catalytic Domains

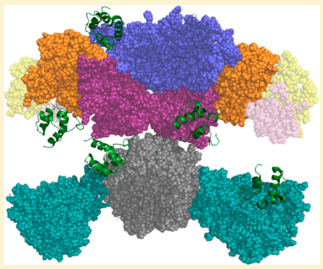
Published as part of The Journal of Physical Chemistry virtual special issue "Manuel Yáñez and Otilia Mó Festschrift".

Matilde F. Viegas, Rui P. P. Neves, Maria J. Ramos, and Pedro A. Fernandes*

REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, s/n, Rua do Campo Alegre, 4169-007 Porto, Portugal

Supporting Information

ABSTRACT: Human fatty acid synthase (hFAS) is a megasynthase whose main function is *de novo* biosynthesis of saturated fatty acids. Interest has been drawn to this enzyme beyond its physiological role due to the association between high levels of hFAS and clinical conditions such as obesity, diabetes, and cancer. Thus, it has become an undeniably attractive pharmacological target. Until now, no crystal structure of the complete hFAS is available, hindering attempts to fully understand this protein. Using homology modeling, we built a model of the entire megasynthase, encompassing all of its domains, including the acyl carrier protein (ACP) and thioesterase (TE) mobile domains absent in the crystal structure of mammalian fatty acid synthase (FAS). On a second stage, we used data-driven protein−protein docking between the substrate shuttling domain ACP and every catalytic domain in the protein. We also propose sets of amino acids at the interface of each domain that we believe are important to favor the interaction between ACP and each domain of hFAS. After inspection, we validated each complex between ACP and MAT/KS/KR/DH/ER domains through classical molecular dynamics simulations and RMSd analysis. Additionally, we mapped the interactions between the residues at the active site of each catalytic domain and its intermediaries. In every docking, we ensured that the distance between catalytic residues and the intermediaries was maintained. Until now, there was not a complete 3D model of this megasynthase. This study is the first to present a homology model for the whole hFAS, including its two mobile domains and possible poses of ACP throughout the cycle of fatty acid biosynthesis, thus mapping obligatory checkpoints in its trajectory. Hence, we believe that these structural insights will allow for future studies of the catalytic mechanism of the overall hFAS.



1. INTRODUCTION

Fatty acid synthase's (FAS) main function is *de novo* biosynthesis of saturated fatty acids. Fatty acids intervene in a myriad of biological processes, acting as energy sources, contributing to membrane integrity and to signal transduction.[1] FAS expression in human adults is low in most tissues,[2] as the surplus of fatty acids acquired in a typical western diet demands for almost no *de novo* fatty acid biosynthesis. However, high levels of this enzyme are associated with a number of clinical conditions ranging from obesity to diabetes and, ultimately, cancer.[3−5]

Regardless of the host organism, fatty acid biosynthesis involves a set of conserved chemical reactions. In bacteria and plants, fatty acid synthesis is accomplished by sequential reactions performed by an ensemble of monofunctional free-standing enzymes classified as Type II FAS. Oppositely, in higher eukaryotes such as fungi and mammals, type I FAS megasynthase is employed, compromising all catalytic domains in a single protein. Recently, it became clear that fungi and animals actually produce fatty acids using both systems: the cytosolic is classified as Type I FAS whereas mitochondrial machinery belongs to Type II FAS.[6] The function of the latter is yet to be established but studies refer it as important to the synthesis of 8C acyl chains involved in posttranslational modification of several mitochondrial proteins.[7−9]

Mammalian FAS (mFAS) is a homodimer of 270 kDa polypeptide chains, coiled and oriented head-to-head.[10] Each of the two mFAS monomers contains all seven catalytic domains: malonyl/acyl transferase (MAT), β-ketosynthase (KS), dehydratase (DH), enoylreductase (ER), β-ketoreductase (KR), thioesterase (TE), and acyl-carrier protein (ACP). The last is a very flexible domain responsible for the substrate shuttling across all of the FAS domains.

De novo biosynthesis of fatty acids consists of a primer substrate, acetyl-CoA, being cyclically condensed with malonyl moieties, and assembled on a phosphopantetheinethiol, which

Received: September 30, 2017
Revised: December 4, 2017
Published: December 6, 2017

ACS Publications © 2017 American Chemical Society

DOI: 10.1021/acs.jpcb.7b09645
J. Phys. Chem. B 2018, 122, 77−85

is covalently linked to ACP. In detail, mFAS fatty acid *de novo* biosynthesis begins at the MAT is provided by acetyl-CoA. This acetyl moiety is loaded onto the ACP substrate shuttling domain, covalently binding its prosthetic group, phosphopantetheine (PNS). The primer substrate is carried from MAT to the active site cysteine of KS. Meanwhile, a malonyl group, the unit used for fatty acid chain elongation, is transferred from Malonyl-CoA to ACP also by MAT. KS will then catalyze the decarboxylative condensation of malonyl and acetyl groups resulting in a β-ketoacyl intermediate bound to ACP. Subsequent reduction is performed by sequential action of NAPDH-dependent KR, DH, and NADPH-dependent ER, resulting in a product elongated by 2C by the end of each cycle. After the completion of 7 cycles, a 16C fatty acid is released by the chain-terminating TE domain.[11]

In 2008, a long-sought crystal structure of mFAS was obtained; at 3.2 Å resolution, it covers five catalytic domains. The flexibly tethered terminal acyl carrier protein and thioesterase domains were absent due to technical issues.[12] Nowadays, homology modeling allows prediction of the structure of a protein using known structures of homologous proteins as templates. This method stems from the principle that the three-dimensional structure of a protein is far more conserved than its primary sequence.[13] Using the crystal structure of mFAS as a framework, we modeled the complete human FAS model; this includes the aforementioned absent ACP and TE domains, as well as missing residues located at loop regions. Intending to further study the works of hFAS, we used HADDOCK[14,15] to perform molecular dockings between ACP and the remaining five domains.

Resulting complexes were validated through conventional molecular dynamics (cMD) simulations.

Combining a new model for the human FAS and molecular docking assays, our results shed light on the overall dynamics of such an unusual enzyme, i.e., hFAS megasynthase.

2. METHODS

2.1. Modeling Human FAS. The human FAS sequence (chain A, 2511 residues) was obtained from the Uniprot database {http://www.uniprot.org}. The sequence was used as a target for homology modeling of the human FAS, all of which was performed by the Swiss-Model Web server. The crystal structure of mammalian *sus scrofa* FAS, PDB code 2vz9, was used as the structural template for residues 1−2113 of both chains A and B, identity percentage 78%, similarity 0.54) and four models were proposed by Swiss-Model server. Global model quality estimation (GMQE) was used to identify the best model from the predicted initial structures.

The obtained model lacked both ACP and TE mobile domains (which are not present in the crystal structure used as template, PDB code 2vz9). To evaluate our model, we performed cMD simulations of *sus scrofa* FAS (template) and hFAS (model), independently. Evolution of backbone RMSd (Figure 1B) shows very similar behaviors between the two; however, the cMD revealed that the loops present in the ψME domain of our model lost their secondary structure (Figure 1C). To improve the quality of homology modeling of the denaturating domain, we introduced the sequence of the ψME domain (1121−1405) in Swiss-Model Web server to find a new template. The crystal structure of Psi/KR Tri-Domain (containing ψME, ψKR, and KR domains), PDB code 4PIV (100% identity, similarity 0.59), presented the 100% identity; therefore, it was used to replace the former domain in our model. A new cMD was performed and the ψME domain no longer caused abnormal alterations in the overall RMSd values of the model, and its secondary structure was preserved (Figure 1D). We proceeded to add the ACP and TE mobile domains, which were still absent at this point. From the complete sequence of human FAS, we identified ACP and TE domains as 75 amino acids (~2120−2196) for ACP and 285 amino acids for TE (~2215−2500). We searched for templates of these two sequences in Swiss-Model and retrieved PDB code 2CG5 as the template for ACP and PDB code 2PX6 as the template for TE. Both crystal structures are of human FAS; thus, no modeling was required. Linkers connecting KR to ACP (Lys2114−Ser2123, 10 residues long), and ACP to TE (Asp2195−Arg2220, 26 residues long) domains, were built from hFAS FASTA sequence using X-Leap. Once the model was complete and fully assembled—total of 5000 residues, without ligands—it was subjected to cMD (details further below).

2.2. Docking Models and Procedures. With a complete model of human FAS, we aimed to build models of the interaction between ACP and the substrate shuttling domain and the remaining five domains using protein−protein docking.

The ACP domain was directly retrieved from PDB code 2CG5 (residues 2125−2194). Five truncated proteins were created from our hFAS model, one for each domain (excluding ACP and TE domains): MAT [415−820], KS [1−406], KR [1879−2113], DH [858−1103], ER [1527−1863] (residue numbers according to our model, PDB file in the Supporting Information).

The HADDOCK Web server was used for protein−protein docking. To do so, a set of active residues needs to be defined. These residues are solvent accessible and thought to be involved in the interaction between the two molecules. For ACP, the set of active residues was established according to the hydrophobic patches described by Bunkoczi et al.:[18] 2137Ile, 2143Val, 2152Leu, 2176Val, 2157Leu, 2158Met (numbering according to PDB code 2CG5). These hydrophobic patches interact with hydrophobic residues at the surface of the partner protein. Thus, for the remaining five domains, active residues were selected on the basis of (i) their presence at the surface of the protein; (ii) their proximity to the active site and (iii) their hydrophobic character.

The resulting sets of active residues were Phe671, Ala769, Val771 (MAT); Leu203, Val261, Phe263 (KS); Leu1979, Ala2029, Ile2079 (KR); Val953, 882Gly, 924Ile (DH); and Gly1590, Ala1619, Phe1632 (ER). Results from each job were ranked according to the HADDOCK score, which is indicative of docking quality. However, HADDOCK did not allow modified residues, such as PNS and substrates, to be present in the submitted PDB. This prohibited us from introducing additional constraints to conduct docking. To overcome this, we had to evaluate the quality of each of the obtained complexes taking in consideration additional constraints resulting in a workflow pipeline:

(i) The ACP domain docked at each of the five domains must not overlap any region of the complete hFAS 3D model (evaluated with Pymol).

(ii) The positioning of Ser2156, to which PNS prosthetic group is bound, must allow the PNS to enter the active site.

Complexes that obey these constraints were subjected to cMD. To do this, we included conserved serine-bound PNS (on ACP) and NADP+ on reductase domains, as well as

DOI: 10.1021/acs.jpcb.7b09645
J. Phys. Chem. B 2018, 122, 77−85

substrates in each domain. Parameters for substrates and serine−PNS group (PNS was coupled with the conserved Serine2156, forming a single modified residue) were obtained with the Antechamber tool, as implemented in AMBER12, using the GAFF force field. Atomic charges were derived using RESP methodology at the HF/6-31Gd level of theory. The NADP+ cofactor parameters were those determined by Ryde.[16] All parameters are included in the Supporting Information.

Further analysis and validation of the complexes were performed by cMD (detailed below).

2.3. Molecular Dynamics Simulations. To assess model and complex stability, structures were subjected to cMD. Prior to simulations, hFAS was parametrized with the ff99SB force-field, and hydrogen atoms and counterions (76 Na+) were added to the final hFAS model using X-Leap, to neutralize it. A total of 445 783 TIP3P water molecules were added within 12.0 Å from the modeled system, to fill a rectangular box. cMD simulations were performed in AMBER12. Minimization was performed in four stages: in the first one, only water molecules were minimized; in the second one, hydrogen atoms were minimized; in the third one, side chains were minimized; and in the last one, all of the atoms in the system were allowed to move. After the minimization procedure, we heated the system from 0 to 310 K and performed a *NVT* 80 ps long simulation maintaining a constant volume, and using periodic boundary conditions. Then, a preparatory production phase was conducted for 2 ns, under periodic boundary conditions with the isobaric−isothermal (*NPT*) ensemble defining a pressure of 1 atm and a temperature of 310 K, using the Langevin thermostat and the Berendsen barostat. The cutoff for Lennard−Jones interactions was set to 9 Å and the Coulomb interactions were treated using the particle mesh Ewald (PME) method, with a cutoff of 9 Å, as well. The time step of the simulation was 2 fs, and the SHAKE algorithm[17] was employed to constrain hydrogen covalent bonds. The propagation of the equations of motion was done with the LeapFrog integration algorithm. The following production run was performed in the same conditions as the preparatory production phase with exception of the cutoff, which was set to 10 Å. The time of the production cMD is variable, ranging from 20 to 24 ns, as stated in the results. Concerning the cMd simulations, in which additional constraints were applied to maintain the catalytic distance for each reaction, the force constant was 50 (kcal/mol)/Å^2. The atoms to which these constraints were applied are discriminated in the results section. In the ER:ACP docking, harmonic restraints were applied to backbone atoms (CA, C, N, O), with a force Constant of 50 (kcal/mol)/Å^2, for the initial 4 ns of the cMD.

3. RESULTS AND DISCUSSION

Formation of protein−protein complexes is a key step in many biological reactions. For fatty acid *de novo* biosynthesis, ACP transient interaction with partner domains is required. A clear perspective on ACP interaction with remaining hFAS domains is unequivocally relevant for drug design directed at health conditions such as obesity, diabetes, and cancer. The current absence of a complete 3D model of human FAS, with ACP bound into the several catalytic domains, has hindered the improvement of drug design studies related to this enzyme, as the competitive inhibitors to be found need to displace the protein−protein complexes between ACP and the other FAS catalytic domains.

In this section, we first present and discuss results from homology modeling performed for hFAS and, on a second stage, the dockings performed on HADDOCK between ACP and five catalytic domains (MAT, KS, KR, DH, and ER).

3.1. Human FAS Model. The complete experimental crystal structure of hFAS is yet to be solved. We propose a homology modeling model accomplished with Swiss-Model homology modeling Web server, using 2vz9 (*sus scrofa* crystal structure of FAS) as the template (∼78% identity and 0.54 similarity). This hFAS structure lacked ACP and TE domains, as well as its connecting linkers (Figure 1A). Twenty-four nanosecond cMD of the homology-modeled hFAS and crystal structure mFAS show that the behaviors of both are similar (Figure 1B), validating our structure as viable.

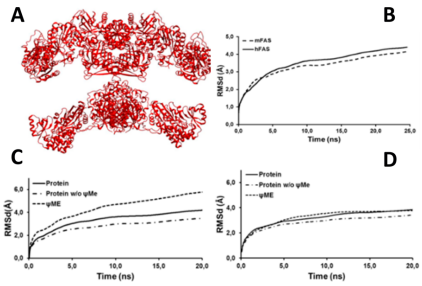


Figure 1. Homology model of human FAS. (A) Human FAS model resulting from Swiss-Model Server, without remodeling of the ΨME domain. (B) Time-dependent root-mean-square deviation (RMSd) (Å) plot of 24 ns cMD comparing the crystal structure of *sus scrofa* FAS (mFAS) and our model of human FAS (hFAS). (C) hFAS model before remodeling the ΨME domain. The ΨME domain behavior is unstable, as it loses its secondary structure throughout the simulation. (D) 20n cMD of hFAS model after remodeling the ΨME domain. This domain displays now very similar values to the ones of the whole complex. Its secondary structure and folding are maintained.

Surveying each of hFAS domains' behavior, loops at the noncatalytic domain ΨME displayed erratic motion losing their secondary structure during the cMD. We remodeled ΨME with the human Psi/KR X-ray as a template, the root mean-square deviation (RMSd) values of the ΨME domain reached a plateau at 4 Å around the 10th ns of simulation (Figure 1C or Figure S1A,B). In addition, an inspection of the secondary structure of ΨME showed no change in the overall fold. The RMSd of the overall hFAS is also approximately independent of the inclusion of the ΨME during the calculation. Therefore, we proceeded with the modeling of the ACP and TE domains, as well as their linkers. The resulting structure displays no abnormally high RMSd regarding the ΨME domain (Figure 1D).

The final model of hFAS was assembled (Figure 2A), connecting the portion not resolved in the original *sus scrofa* X-ray structure [linker_ACP_linker_TE] to the appropriate residue in the KR domain (in both monomers). The resulting solvated system, of 1.5 million atoms, was subjected to 20 ns cMD simulation. Given the dimension of the megasynthase, its overall RMSd goes up to 12 Å (Figure 2B). Visual inspection of the cMD simulation and superposition of the initial and final structures clarified the reason behind this: a tilting motion between the top and bottom section of the enzyme, as well as

DOI: 10.1021/acs.jpcb.7b09645
J. Phys. Chem. B 2018, 122, 77−85

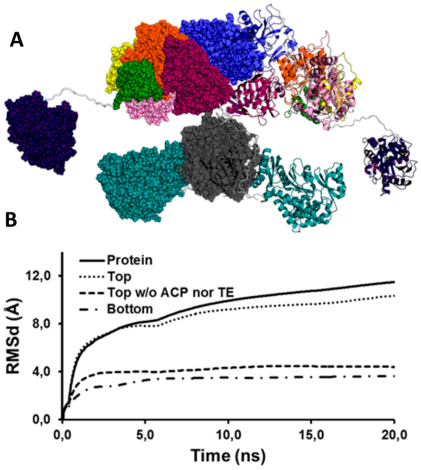


Figure 2. Homology model of human FAS. (A) Human FAS complete homology model. Chain A is displayed in van der Waals representation, and chain B is in cartoon representation: MAT (teal), KS (gray), DH (magenta), ER (blue), KR (orange), ΨME (yellow), ΨKR (pink), ACP (green), TE (dark blue). Linkers are shown in white. (B) Time-dependent root-mean-square deviation (RMSd) (Å) plot for 20 ns cMD simulation of the hFAS complete model: protein (complete line); top section of the protein consisting of domains DH, ER, KR, ΨME, ΨKR, ACP, TE (dotted line); top section of the protein without ACP and TE mobile domains (dashed line); bottom section of the protein consisting of MAT and KS domains (dash/dot line).

the intrinsic mobility of ACP and TE domains (Figure 2B). When the RMSd values are calculated for each domain, the profile is cohesive and stable; all domains display RMSd below 4 Å (Figure S2), and their folding is maintained throughout the simulation.

3.2. Docking of ACP to the Catalytic Domains. With a final refined model to work with, we studied the interaction between ACP and the remaining domains of hFAS. As previously stated, ACP is responsible for substrate shuttling across all catalytic domains, in a fashion that is yet to be understood. In type I FAS, ACP is attached to the KR domain through a linker and binds substrates through a sulfur atom at the tip of its prosthetic group phosphopantetheine, PNS.

Posttranslational modification of ACP conserved serine with a phosphopantetheinyl moiety is carried out by phosphopantetheinyltransferase (PPT), resulting in holo-ACP. A study of the mechanism and substrate recognition of *Homo sapiens* ACP has brought to light the nature of interaction between hFAS, ACP and PPT: its binding is unusual and mostly facilitated by hydrophobic interactions between hydrophobic patches present at both ACP and PPT surfaces.[18] On the basis of this study, we first searched for hydrophobic residues neighboring the entrance to the active site of each of the five domains, restricting the selection to at least three residues to conduct each docking. Considering these and ACP's hydrophobic patches, we designated the set of active residues, as reported in the Methods section, to perform HADDOCK protein−protein docking assays in which ACP is docked at each one of the five domains of hFAS (Figure 3).

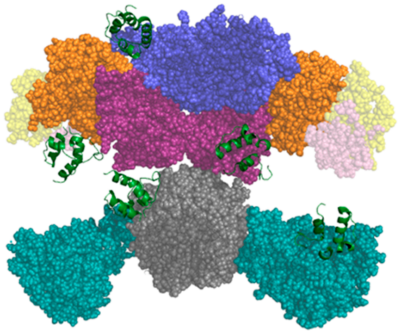


Figure 3. ACP domain docked at MAT, KS, KR, DH, and ER catalytic domains. ACP is displayed in green cartoon. The hFAS main body is displayed in van der Waals representation: MAT (teal), KS (gray), DH (magenta), ER (blue), KR (orange). Noncatalytic domains, ΨME (yellow) and ΨKR (pink), are also displayed in van der Waals representation. For clarity, the TE domain, and its connecting linkers, are absent in the picture.

Further validation should be conducted for our models to be of use for future drug design campaigns such as virtual screening. To do that, we would need to study the hydrophobic lining of the active site pocket of each of the hFAS domains, modeling an elongated form of the fatty acid chain, instead of the shorter form of the substrates. Nevertheless, we analyze the main interactions established between the interfaces of ACP and each of the five catalytic domains, as well as the main interaction in the catalytic pocket of each of these domains. Hence, our study should be seen as starting point for future *in silico* drug design campaigns.

3.2.1. MAT:ACP. In Type I human FAS, the MAT domain is responsible for the initiation of *de novo* fatty acid biosynthesis. It is responsible for loading the primer substrate, an acetyl moiety, from acetyl-CoA into the ACP domain and, given its bifunctional characteristics; it is also responsible for loading malonyl, the extending unit, into ACP in the following cycles until the finishing point.

Malonyl-CoA is the substrate required for elongation of fatty acid chain. A malonyl moiety will result from the cleavage of the substrate and will bind the conserved serine at the MAT active site. Then, the ACP's PNS sulfur will bind the carboxyl carbon of malonyl available at the MAT active site, transferring it to the ACP domain. A mechanistic study on MAT catalysis developed in our group (under submission work from our group) has provided us with the information to determine the necessary atomic environment for both reactions to occur. In our assays, we focused on the second step of the reaction and our starting structures reflected these conditions: the ACP domain docked at the MAT domain, onto which a malonyl moiety is already bonded (Figure 4D).

To conduct HADDOCK docking assays between ACP and MAT domains, we selected active residues. For the ACP domain, the previously defined set (Methods) was selected. MAT residues were chosen according to the mapping of the exposed region enclosing the binding pocket.[19] Furthermore,

DOI: 10.1021/acs.jpcb.7b09645
J. Phys. Chem. B 2018, 122, 77−85

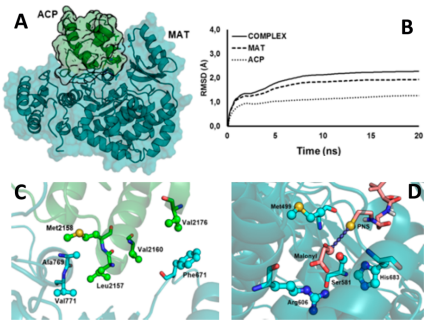


Figure 4. ACP docked at MAT. (A) ACP is displayed in green cartoon and surface, and MAT is in blue cartoon and surface. (B) Time-dependent root-mean-square deviation (RMSd) (Å) plot for the MAT:ACP complex, during a 20 ns restraint-free cMD simulation. The RMSd trajectories of the complex are shown in black full line, MAT RMSd values in dashed line, and ACP RMSd values in dots. (C) Residues selected as active for conducting HADDOCK assays. MAT residues are in blue (Phe257, Ala355, Val356), and ACP residues are in green (Leu2157, Met2158, Val2160, Val2176). Residue side chains are shown in ball and stick representation. (D) PNS (phosphopantetheine) and malonyl (substrate) are shown in coral stick representation, and MAT active site residues are in blue (Met85, Ser167, Arg192, His269). The dashed line shows the distance restriction of 4.0 Å applied in the first 8 ns of the cMD simulation (for details see Figure S3).

hydrophobic residues were preferred to complement the binding hydrophobic patches of ACP.[18] We chose Phe671, Ala769, and Val771 on the surface of MAT domain (Figure 4C). After the HADDOCK run, we selected a set of six complex structures on the basis of visual inspection of complementarity of domains ACP and MAT, and on the basis of the distances between their active residues. These six complexes were subjected to a cMD, after which, only one did not dissociate during the 20 ns cMD run.

We also wanted to evaluate if the remaining complex could reproduce the conserved interactions between the substrate and the MAT catalytic domain. To do so, we followed the interactions of PNS and malonyl with the active site of MAT during a 20 ns cMD. After pulling out the restraints in phosphopantetheine's sulfur and malonyl's C1, which are involved in the malonyl moiety transfer from MAT to ACP, these remained 3.84 ± 0.53 Å apart (consistent with a reactive distance Figure S3C). Just as reported in Maier et al.,[12] the MAT domain has a conserved Ser−His dyad, a conserved Arg for salt-bridging malonyl substrates, and Met that contributes to a hydrophobic pocked. In our model, these residues identify as His683, Ser581, Arg 606, and Met499 (Figure 4D).

Given the importance of these residues, we set to study their relative positions. The distance between the Ser581 and His683 stabilized at 3.7 ± 0.49 Å. The Met499, important for the maintenance of the hydrophobic character of the pocket, was 3.06 ± 0.24 Å distant from the substrate (malonyl). Finally, the Arg606 was 4.06 ± 0.15 Å away from the substrate as well. Thus, we can infer that the model was capable of sustaining the active site residues disposition throughout the 20 ns cMD. Altogether, the selected pose remained stable during the cMD simulation (RMSd values below 2 Å, over 20 ns, Figure 4B).

Regarding the interface between ACP and MAT, we focused our attention on the selected active residues form each domain (Figure 4C). We expected to see consecutive hydrophobic interactions forming hydrophobic patches. The β-carbon of Ala769 interacts with the CG of Leu2157 (6.03 ± 1.34 Å) and CG of Met2158 (4.37 ± 0.52 Å). However, this Met2158, interacted with Ile2239 (4.41 ± 0.60 Å), both present in the ACP. These four hydrophobic residues add up their volume forming one patch. Another set of residues includes Phe671 which interacts with Val2160 (5.91 ± 0.89 Å) and Val2176 (6.02 ± 1.01 Å), forming a second patch.

3.2.2. KS:ACP. The KS domain not only is structurally relevant but also is a critically important catalytic domain. The structural stability of the mFAS megasynthase depends on the dimeric N-terminal KS domains, which, if absent, causes mFAS to become entirely monomeric.[20] Catalytically, it covalently joins acetate and malonate together, as well as the malonyl unit to the growing fatty acid chain. Catalysis of this reaction depends on the close interaction between the active-site Cys161[19] of KS and the phosphopantetheine moiety of ACP—first, in the transfer of a saturated acyl moiety from phosphopantetheine to cysteine and, second, in the condensation of *S*-acyl cysteine and *S*-malonyl phosphopantetheine moieties. Two other histidine residues in the active site are responsible for stabilizing the intermediate.

To perform ACP:KS docking, we followed the same procedure as with MAT:ACP, defining two sets of active residues to conduct HADDOCK docking. Therefore, we surveyed the KS surface, near the entrance of the active site, for hydrophobic residues. A first selection included residues Ile2137, Val2143, and Leu2152; however, only one structure displayed clear complementarity between the two domains. This structure was obtained with KS active residues Leu203, Val261, and Phe263, located at the surface of KS (Figure 5C).

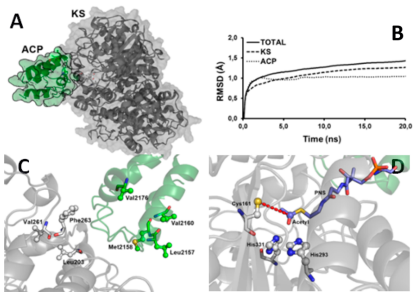


Figure 5. ACP domain docked at KS. (A) ACP is displayed in green cartoon and surface, and KS is in orange cartoon and surface. (B) Time-dependent root-mean-square deviation (RMSd) (Å) plot for the KS:ACP complex, during a 20 ns, no restraints applied, cMD simulation. The RMSd trajectories of the complex are shown in black full line, MAT RMSd values in dashed line, and ACP RMSd values in dots. (C) Residues selected for conducting HADDOCK assays. KS residues are in gray (Leu203, Val261, Phe263), and ACP residues are in green (Leu2157, Met2158, Val2160, Val2176). Residue side chains are shown in ball and stick representation. (D) PNS (phosphopantetheine) and acetyl (substrate) are shown in blue stick representation, and KS active site residues are in gray (Cys161, His293, His331). The dashed line shows the 5 Å distance restriction applied in the first 20 ns of the cMD simulation (for details see Figure S4).

DOI: 10.1021/acs.jpcb.7b09645
J. Phys. Chem. B 2018, 122, 77−85

When subjected to a 20 ns cMD, the hydrophobic interactions between the two domains fluctuated during the cMD due to the abundance of hydrophobic residues at their surface. The ones that stood out were Leu203 and Leu841 (8.37 ± 0.55 Å), Pro302 and Leu846 (7.82 ± 0.92 Å), and Val534 and Ala831 (8.10 ± 1.34). To further evaluate our model, we focused our attention on the active site. During the KS catalytic cycle, Cys161's sulfur performs a nucleophilic attack in the acyl-ACP's carbonyl carbon. The distance between these two atoms must not exceed 5 Å to guarantee a catalytic conformation (unpublished work from our group). Concerning our structure, the distance between these key atoms fluctuated during an initial cMD simulation. So, to emulate the reaction environment, we kept these atoms at the desired distance during 4 ns. Upon removal of these restraints, this distance was maintained (4.16 ± 0.47 Å) conserving the distance for the nucleophilic attack (Figure 5D). The His 293 and His331, part of the active site triad,[19] were at 4.30 ± 0.73 and 4.11 ± 0.32 Å, respectively. Regarding the whole model, it showed an overall backbone RMSd below 1.5 Å (Figure 5D).

3.2.3. KR:ACP. KR domain is responsible for the reduction of β-ketoacyl (KAC) through (i) a nucleophilic attack of a hydride from NADPH to the β-carbon of the substrate and (ii) a deprotonation of an acid by the alkoxide moiety of the β-carbon, to hold the final alcohol product.[21]

To study the interactions between ACP and KR, we followed the previously mentioned protocol. The residues chosen from KR domain were Leu1979, Ala2029, and Ile2079 (Figure 6C). From the HADDOCK run, one of the complexes displayed higher proximity between all the selected active residues, thus showing higher complementarity between the two domains.

Regarding the complex, we can see that the RMSd does not reach a plateau after 13 ns, a result of a tilting motion between ACP and the KR domain. None of the domains denatured, according to the RMSd of each domains. This tilting motion occurred as one of the hydrogen bonds weakened dramatically (Leu1979CG and Val2170CB, 21.60A ± 5.67 Å). Still, four other hydrogen bonds were stable throughout the cMD, holding the complex together. The four hydrogen bonds are Met2073SD and Met2152SD (6.70 ± 1.69 Å), Val2080CG and Met2152.CE (7.20 ± 1.39 Å), Ile2079CG and Met2152.CE (9.93 ± 1.59 Å), and Ala2029CB and Leu2151CD1 (8.75 ± 2.62 Å).

Regarding the active site, the catalysis-prone environment was ensured through restraints on the substrate's oxygen and tyrosine OH group (Figure 5D). After removal of the restraints, we observed that the distance between these two atoms was stable at 1.76 ± 0.08 Å, as well as between NDP238.O2D and KAC303.C at 3.17 ± 0.67 Å, ensuring the adequacy of the model for catalysis.

3.2.4. DH:ACP. DH domain is responsible for removal of the hydroxyl group from the β-OH group of the ACP-bond substrate β-hydroxylacetyl (HAC). Its catalytic dyad, His878, Asp1031, is highly conserved across species.[22] Mutagenesis of Gln1035, present in the immediacy of the His and Asp residues, led to reduction of DH enzymatic activity. This residue is responsible for anchoring the position of the catalyti1c aspartate Asp 1031.[22] Structurally, the active site residue His878, whose mutation eliminates all DH activity,[23] is present at the DH1 while Asp1031 and Gln1035 are present at DH2.[24,25]

To conduct the docking assay, we selected residues Gly882, Ile924, and Val953, present at the surface of DH, as active residues. The usual Leu2157, Val2160, and Val2176 residues on ACP surface were provided as active residues (Figure 7C).

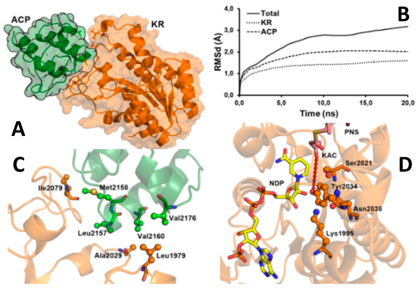


Figure 6. ACP domain docked at KR. (A) ACP is displayed in green cartoon and surface, and KR is in orange cartoon and surface. (B) Residues selected for conducting HADDOCK assays. KR residues are in orange (Leu1979, Ala2029, Ile2079), and ACP residues are in green (Leu2157, Met2158, Val2160, Val2176). Residue side chains are shown in ball and stick representation. (C) Time-dependent root-mean-square deviation (RMSd) (Å) plot for the KR:ACP complex, during a 20 ns cMD simulation without restraints. The RMSd trajectories of the complex are shown in black full line, KR values in dashed line, and ACP RMSd values in dots. (D) PNS (phosphopantetheine) and KAC (substrate) are shown in coral stick representation, and KR active site residues are in orange (Lys1995, Ser2021, Tyr2034, Asn2038). The red dashed line shows the 1.8 Å distance restriction applied in the first 4 ns of the cMD simulation (for details see Figure S5).

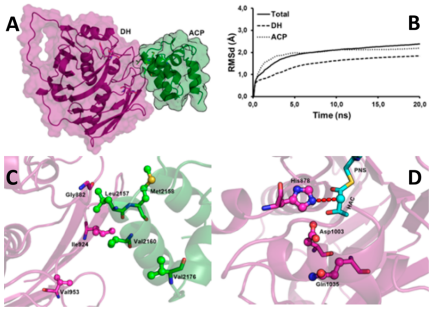


Figure 7. ACP domain docked at DH. (A) ACP is displayed in green cartoon and surface, and DH is in purple cartoon and surface. (B) Time-dependent root-mean-square deviation (RMSd) (Å) plot for the DH:ACP complex, during a 20 ns cMD simulation. The RMSd trajectories of the complex are shown in black full line, DH RMSd values in dashed line, and ACP RMSd values in dots. (C) Residues selected for conducting docking. DH residues are in purple (His878, Asp1003, Gln1035), and ACP residues are in green (Leu2157, Met2158, Val2160, Val2176). Residue side chains are shown in ball and stick representation. (D) PNS (phosphopantetheine) and HAC (substrate) are shown in cyan stick representation, and DH active site residues are in purple (His878, Asp1003, Gln1035). The red dashed line shows the 4.2 Å distance restriction applied in the 4−8 ns time window of the cMD simulation (for details see Figure S6).

DOI: 10.1021/acs.jpcb.7b09645
J. Phys. Chem. B 2018, 122, 77−85

The selected complex showed proximity between active residues Gly882 and Leu2157 (8.26 ± 0.64 Å) and Ile924 and Val2160 (9.36 ± 0.44 Å). Ile924 established an interaction with Val884 (5.56 ± 0.38 Å) and with Leu885 (4.97 ± 0.91 Å), forming a patch. However, Leu2157 from ACP interacted with Met2158 (4.57 ± 0.71 Å). The residue Val 953 did not display interactions with other partnering residues on the ACP surface. The cMD simulation of the chosen complex reveals a stable looking complex, as seen in Figure 7B RMSd plot. Looking at the active site, HAC and the catalytic His878[22] maintain a 4.53 ± 0.30 Å distance throughout the cMD. Simultaneously, Asp1033 and Gln1035 formed a hydrogen bond (4.10 ± 1.75 Å) and the Ala888 helped to keep the product in place by another hydrogen bond (2.18 ± 0.32 Å) (Figure 7D).

3.2.5. ER:ACP. The ER domain is involved in the last step of the fatty acid chain elongation, catalyzing the reduction of *trans*-2-unsaturated fatty acids. Just like the KR domain, ER catalysis is dependent on NADPH as a cofactor for hydride transfer. Thus, the ER active site binds not only the substrate but also the cofactor through hydrogen bonds.

The crystal structure of the human ER has been solved, both in the presence and in the absence of an inhibitor, triclosan (TCL). However, this did not disclose any information about the mechanism of this domain: TCL is an allosteric inhibitor whose binding site is located at the protein−protein interface, causing downstream structural rearrangements affecting critical active site residues.[26] To map out the active site, we matched several facts about the ER domain. First, we know where the NAPDH is located in the domain because of the X-ray structure of *sus scrofa* (PDB code 2vz9),[12] from Val1651 to Val1794, but the substrate position is still obscure. Additionally, prior work on the Type I KR domain, produced at our laboratory, explained in detail the nature of the reductase reaction.[21] Work on the Type I FAS ER domain has proposed the Lys1771 and Asp1797 residues, present at the active site, as important residues for the cofactor binding;[27] thus, they present themselves as active site residues candidates. To unveil the position of the reactant in the active site, we searched for structurally similar proteins. We found an X-ray structure of *Arabidopsis thaliana* reductase (PDB code 2J3K) whose structure is highly similar to that of ER.[28] Using this as template, we forayed into positioning a substrate in our model. On the basis of the surface of the domain, as well as proximity to an access tunnel to NAPDH, we chose two sets of hydrophobic active residues to perform the HADDOCK run: [Gly1590, Ala1619, Phe1632] vs [Ala1619, Leu1627, Phe1632]. The best structure was result of the former set (Figure 8A,C); we chose only one structure as none of the others displayed poses that allowed for the PNS prosthetic group to be inserted into the active site.

Considering that, in the previous section, we detailed the TBT product of DH catalysis, we opted for using it as a substrate in the ER domain. Due to the lack of information on the active site, and which of the active residues we should apply restraints to, we opted for fixing only TBT position, during the 20 ns.

During the 20 ns cMD, we observed the interactions of the active residues selected for the HADDOCK run. Not all of the initial set of active residues intervened in the ACP:ER interaction. The most relevant hydrophobic patches were (i) Leu373 and Met274 (5.74 ± 0.72 Å) and Leu373 and Phe104 (7.63 ± 1.36 Å), while Leu373 also interacts with Val376; (ii) Pro28 and Leu368 (5.98 ± 1.08 Å), where Leu368 interacts with Leu400 and Val359, while Pro28 interacts with Leu23 and Ala26; (iii) Pro102 and Met374 (7.06 ± 1.63 Å), where Pro102 interacts with Phe102 and Leu105, while Met374 binds to Leu373 and Ile350.

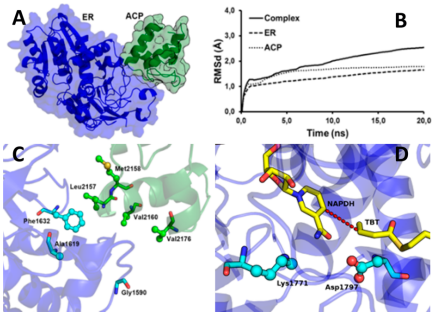


Figure 8. ACP domain docked at ER. (A) ACP is displayed in green cartoon and surface, and ER is in blue cartoon and surface. (B) Residues selected for conducting docking. ER residues are in cyan (Gly1590, Ala1619, Phe1632), and ACP residues are in green (Leu2157, Met2158, Val2160, Val2176). Residue side chains are shown in ball and stick representation. (C) Time-dependent root-mean-square deviation (RMSd) (Å) plot for the ER:ACP complex, during a 32 ns cMD simulation. The RMSd trajectories of the complex are shown in black full line, ER RMSd values in dashed line, and ACP RMSd values in dots. (D) PNS (phosphopantetheine) and TBT (substrate) are shown in blue stick representation, and ER active site residues are in cyan (Tyr1771, Lys1797). The red dashed line shows the 5.9 Å distance between NAPDH and TBT (for details see Figure S7).

As for the active site (Figure 7D), we first looked at the distance between TBT and reduction agent NADPH; it was 5.62 ± 1.51A. The PNS prosthetic arm is stabilized by hydrogen bonds; however, in the pocket where the substrate is located, we noticed that the residues lining the pocket are hydrophobic (Phe, Met, Leu). This concurs with the fact that the ER active site pocket is probably the second most hydrophobic pocket in the hFAS megasynthase because the substrate possesses a carbon chain devoid of polar atoms. This environment reduces the number of neighboring acidic residues in the pocket. From previous work done in our group, we know that the KR domain reaction involves a reducing agent, NAPDH, and an acid required for the stabilization of the carbanion formed in the first step. Given the hydrophobicity of the pocket, the closest acidic residue to TBT is Lys243. In our cMD, the Lys starts with an orientation incompatible with its role of stabilizing the carbanion; however, as the cMD progresses, it reorients (8.38 ± 1.73 Å). The vehicle for the transfer of a proton is still unknown, however, we can speculate it is either through a water molecule or a through the ribose, just like in KR. Also, the Asp47 is very close to the Lys243, which might indicate its role in stabilizing the Lys243. The distance between Lys243 and Asp47 is 6.71 ± 1.58 Å.

4. CONCLUSION

This is an unprecedented study using a 1.5 million atoms model of a highly complex fatty acid megasynthase, hFAS. Packing 14 catalytic domains in two monomers, this is a challenging protein to obtain a X-ray crystallographic structure, delaying

DOI: 10.1021/acs.jpcb.7b09645
J. Phys. Chem. B 2018, 122, 77−85

possible scientific breakthroughs. Using *in silico* approaches, such as homology modeling, docking, and MD simulation, we provide a refined and highly plausible structure of the human fatty acid synthase.

On a second stage of the project, dockings were performed between ACP substrate shuttling domain and five other catalytic domains, providing clear insight on the interaction between domains, as well as an overview of their active site. This also proved the viability of the structure, proving its utility in interactions studies such as this one.

Despite the growing extent of available literature on this enzyme, no other work has dwelt with enzyme in such depth. Our homology model and *in silico* studies not only give insight into such a mysterious megasynthase but also provide the scaffold for future human FAS inhibition studies, useful for pharmacologic research on cancer, diabetes and obesity.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcb.7b09645.

Detailed RMSd plots for each of the cMD mentioned in the paper, as well as distance plots between selected residues (PDF)

Coordinate files for the newly constructed model (PDB) (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*P. A. Fernandes. E-mail: pafernan@fc.up.pt.

ORCID

Matilde F. Viegas: 0000-0002-4472-9691

Rui P. P. Neves: 0000-0003-2032-9308

Maria J. Ramos: 0000-0002-7554-8324

Author Contributions

M. F. Viegas performed all the calculations, P. P. R. Neves helped in supervising the work, and P. A. Fernandes and M. J. Ramos designed the computational experiments and supervised all the work. The manuscript was written through contributions of M. F. Viegas, P. A. Fernandes, and M. J. Ramos. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work received financial support from the following institutions: Fundação para a Ciência e a Tecnologia (FCT) through project EXCL/QEQ-COM/0394/2012 and EXCL-II/QEQ-COM/0394/2012.

■ ABBREVIATIONS

ACP, acyl carrier protein; cMD, classical molecular dynamics; ER, enoylreductase; Fas, fatty acid synthase; GMQE, global model quality estimation; DH, dehydratase; HAC, substrate β-hydroxylacetyl; hFAS, human fatty acid synthase; KAC, β-ketoacyl; KR, β-ketoreductase; KS, β-ketosynthase; MAT, malonyl/acyl transferase; mFAS, mammalian FAS; NADPH, nicotinamide adenine dinucleotide phosphate; PME, particle mesh Ewald; PNS, phosphopantetheine; PPT, phosphopantetheinyl transferase; RMSd, root-mean-square deviation; TBT, *trans*-2-butyryl; TCL, triclosan; TE, thioesterase

■ REFERENCES

(1) Wakil, S. J.; Abu-Elheiga, L. A. Fatty Acid Metabolism: Target for Metabolic Syndrome. *J. Lipid Res.* **2009**, *50* (Supplement), S138−S143.

(2) Kusakabe, T.; Maeda, M.; Hoshi, N.; Sugino, T.; Watanabe, K.; Fukuda, T.; Suzuki, T. Fatty Acid Synthase Is Expressed Mainly in Adult Hormone-Sensitive Cells or Cells with High Lipid Metabolism and in Proliferating Fetal Cells. *J. Histochem. Cytochem.* **2000**, *48* (5), 613−622.

(3) Shurbaji, M. S.; Kalbfleisch, J. H.; Thurmond, T. S. Immunohistochemical Detection of a Fatty Acid Synthase (OA-519) as a Predictor of Progression of Prostate Cancer. *Hum. Pathol.* **1996**, *27* (9), 917−921.

(4) Gansler, T. S.; Hardman, W.; Hunt, D. A.; Schaffel, S.; Hennigar, R. A. Increased Expression of Fatty Acid Synthase (OA-519) in Ovarian Neoplasms Predicts Shorter Survival. *Hum. Pathol.* **1997**, *28* (6), 686−692.

(5) Rashid, A.; Pizer, E. S.; Moga, M.; Milgraum, L. Z.; Zahurak, M.; Pasternack, G. R.; Kuhajda, F. P.; Hamilton, S. R. Elevated Expression of Fatty Acid Synthase and Fatty Acid Synthetic Activity in Colorectal Neoplasia. *Am. J. Pathol.* **1997**, *150* (1), 201−208.

(6) Bunkoczi, G.; Misquitta, S.; Wu, X.; Lee, W. H.; Kochan, G.; Kavanagh, K. L.; Oppermann, U.; Smith, S. Structural Basis for Different Specificities of Acyltransferases Associated with the Human Cytosolic and Mitochondrial Fatty Acid Synthases. *Chem. Biol.* **2009**, *16* (6), 667−675.

(7) Gueguen, V.; Macherel, D.; Jaquinod, M.; Douce, R.; Bourguignon, J. Fatty Acid and Lipoic Acid Biosynthesis in Higher Plant Mitochondria. *J. Biol. Chem.* **2000**, *275* (7), 5016−5025.

(8) Schweizer, E.; Hofmann, J. Microbial Type I Fatty Acid Synthases (FAS): Major Players in a Network of Cellular FAS Systems. *Am. Soc. Microbiol.* **2004**, *68* (3), 501−517.

(9) Witkowski, A.; Joshi, A. K.; Smith, S. Coupling of the de Novo Fatty Acid Biosynthesis and Lipoylation Pathways in Mammalian Mitochondria. *J. Biol. Chem.* **2007**, *282* (19), 14178−14185.

(10) Staunton, J.; Weissman, K. J. Polyketide Biosynthesis: A Millennium Review. *Nat. Prod. Rep.* **2001**, *18* (4), 380−416.

(11) Smith, S.; Tsai, S.-C. The Type I Fatty Acid and Polyketide Synthases: A Tale of Two Megasynthases. *Nat. Prod. Rep.* **2007**, *24* (5), 1041−1072.

(12) Maier, T. The Crystal Structure of a Mammalian Fatty Acid Synthase. *Science (Washington, DC, U. S.)* **2008**, *321* (2008), 1315.

(13) Martí-Renom, M. a.; Stuart, a C.; Fiser, a.; Sánchez, R.; Melo, F.; Sali, a. Comparative Protein Structure Modeling of Genes and Genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291−325.

(14) van Zundert, G. C. P.; Rodrigues, J. P. G. L. M.; Trellet, M.; Schmitz, C.; Kastritis, P. L.; Karaca, E.; Melquiond, A. S. J.; van Dijk, M.; de Vries, S. J.; Bonvin, A. M. J. J. HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **2016**, *428*, 720.

(15) Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J. HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Data. *J. Am. Chem. Soc.* **2003**, *1737*, 1731−1737.

(16) Holmberg, N.; Ryde, U.; Bülow, L. Redesign of the Coenzyme Specificity in L-Lactate Dehydrogenase from Bacillus Stearothermophilus Using Site-Directed Mutagenesis and Media Engineering. *Protein Eng., Des. Sel.* **1999**, *12* (10), 851−856.

(17) Ryckaert, J.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.

(18) Bunkoczi, G.; Pasta, S.; Joshi, A.; Wu, X.; Kavanagh, K. L.; Smith, S.; Oppermann, U. Mechanism and Substrate Recognition of Human Holo ACP Synthase. *Chem. Biol.* **2007**, *14* (11), 1243−1253.

(19) Pappenberger, G.; Benz, J.; Gsell, B.; Hennig, M.; Ruf, A.; Stihle, M.; Thoma, R.; Rudolph, M. G. Structure of the Human Fatty Acid Synthase KS-MAT Didomain as a Framework for Inhibitor Design. *J. Mol. Biol.* **2010**, *397* (2), 508−519.

DOI: 10.1021/acs.jpcb.7b09645
J. Phys. Chem. B 2018, 122, 77−85

(20) Mullen, G. E.; Yet, L. Progress in the Development of Fatty Acid Synthase Inhibitors as Anticancer Targets. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 4363.
(21) Medina, F. E.; Neves, R. P. P.; Ramos, M. J.; Fernandes, P. A. A QM/MM Study of the Reaction Mechanism of Human [Small Beta]-Ketoacyl Reductase. *Phys. Chem. Chem. Phys.* **2017**, *19* (1), 347−355.
(22) Pasta, S.; Witkowski, A.; Joshi, A. K.; Smith, S. Catalytic Residues Are Shared between Two Pseudosubunits of the Dehydratase Domain of the Animal Fatty Acid Synthase. *Chem. Biol.* **2007**, *14* (12), 1377−1385.
(23) Joshi, A. K.; Smith, S. Construction, Expression, and Characterization of a Mutated Animal Fatty Acid Synthase Deficient in the Dehydrase Function. *J. Biol. Chem.* **1993**, *268* (30), 22508−22513.
(24) Helmicamp, G. M.; Bloch, K. Beta-Hydroxydecanoyl Thioesther Dehydratase Studies on Molecular Structure and Active Site. *J. Biol. Chem.* **1969**, *244* (21), 6014−6023.
(25) Akey, D. L.; Razelun, J. R.; Tehranisa, J.; Sherman, D. H.; Gerwick, W. H.; Smith, J. L. Crystal Structures of Dehydratase Domains from the Curacin Polyketide Biosynthetic Pathway. *Structure* **2010**, *18* (1), 94−105.
(26) Sippel, K. H.; Vyas, N. K.; Zhang, W.; Sankaran, B.; Quiocho, F. A. Crystal Structure of the Human Fatty Acid Synthase Enoyl-Acyl Carrier Protein-Reductase Domain Complexed with Triclosan Reveals Allosteric Protein-Protein Interface Inhibition. *J. Biol. Chem.* **2014**, *289* (48), 33287−33295.
(27) John, A.; Umashankar, V.; Krishnakumar, S.; Deepa, P. R.; Wakil, S.; Menendez, J.; Lupu, R.; Smith, S.; Witkowski, A.; Joshi, A.; et al. Comparative Modeling and Molecular Dynamics Simulation of Substrate Binding in Human Fatty Acid Synthase: Enoyl Reductase and β-Ketoacyl Reductase Catalytic Domains. *Genomics Inform.* **2015**, *13* (1), 15.
(28) Youn, B.; Kim, S.-J.; Moinuddin, S. G. A.; Lee, C.; Bedgar, D. L.; Harper, A. R.; Davin, L. B.; Lewis, N. G.; Kang, C. Mechanistic and Structural Studies of Apoform, Binary, and Ternary Complexes of the Arabidopsis Alkenal Double Bond Reductase At5g16970. *J. Biol. Chem.* **2006**, *281* (52), 40076−40088.

DOI: 10.1021/acs.jpcb.7b09645
J. Phys. Chem. B 2018, 122, 77−85

3.3. Modeling of Type I FAS Thioesterase Domain and preparation of QM/MM model for upcoming catalytic mechanism studies

Matilde F. Viegas[a], Jorge Juan Cabrera-Trujillo[b], Rui P. P. Neves[a], Pedro A. Fernandes[a], and Maria J. Ramos[a]

[a] LAQV@REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

[b] Departamento de Química Orgánica I and Centro de Innovación en Química Avanzada (ORFEO-CINQA), Faculdad de Ciencias Químicas, Universidad Complutense de Madrid, 28040 Madrid, Spain.

The following work is in its early stages. Due to this, this chapter is structured differently to aid the reader. The context for the study is provided by a comprehensive introduction on the TE domain of animal FAS partly compiled and adapted from "Animal Fatty Acid Synthase: a Chemical Nanofactory" [54], followed by the methodology involved in this study as well as the obtained results.

Regarding the contributions to this work, Matilde F. Viegas and Jorge Juan Cabrera-Trujillo performed the calculations with the guidance of Matilde F. Viegas and Rui P. P. Neves, and under the supervision of Pedro A. Fernandes, and of Maria J. Ramos.

3.3.1. Abstract

TE domain is responsible for the termination step of *de novo* fatty acid biosynthesis. After seven to eight elongation cycles, TE hydrolyzes the thioester bond between the long-chain palmitoyl (C16) or stearate (C18) and the 4'-phosphopantetheine moiety of the FAS ACP domain [48,53]. The crucial role of TE is in the determination of the animal FAS product's length, as its activity is closely correlated to the size of its acyl substrate. Animal FAS TE is characterized by a conserved α/β hydrolase fold [231] with two subdomains – A and B. In addition, a loop insertion between strands β5 and β6, named helix flap, is thought to be involved in the TE's substrate selectivity mechanism.
The present work introduces a QM/MM model to study the catalytic mechanism of the human FAS TE domain, with the natural C16 palmitoyl substrate bound to the catalytic Ser2308.

Keywords: Fatty Acid Synthase Type I, Human FAS, Thioesterase, TE, Enzymatic Catalysis, QM/MM, ONIOM

3.3.2. Introduction

TE domain's role in animal FAS

In the context of *de novo* fatty acid biosynthesis, TE catalyzes the termination step after seven to eight elongation cycles, hydrolyzing the thioester bond between the long-chain palmitoyl (C16) or stearate (C18) and the 4'-phosphopantetheine moiety of the FAS ACP domain [48,53].

TE domain is tethered to the animal FAS megasynthase through a flexible linker connected to the KR domain of each FAS polypeptide chains. TE position precedes the ACP domain, and a linker connects these two domains, about 20 residues long, as seen in Figure 6A. Outstandingly, the ACP and TE domains were absent in the X-ray structure of animal FAS [48], suggesting they are remarkably flexible in relation to the core of the megasynthase.

TE's function is crucial to regulating the animal FAS product's length. When isolated from the megasynthase body by limited proteolysis [232], TE's absence affects the length of the resulting fatty acid; without it, the resulting product's profile shifts from 16 or 18 to 20 or 22 carbon-long acyl chains. When fully functioning and attached to the megasynthase, TE's activity is tightly correlated to the length of its acyl substrate. Its activity peaks when the substrate is C16 and lowers for chains longer than C18 or shorter than C16 [232–234].

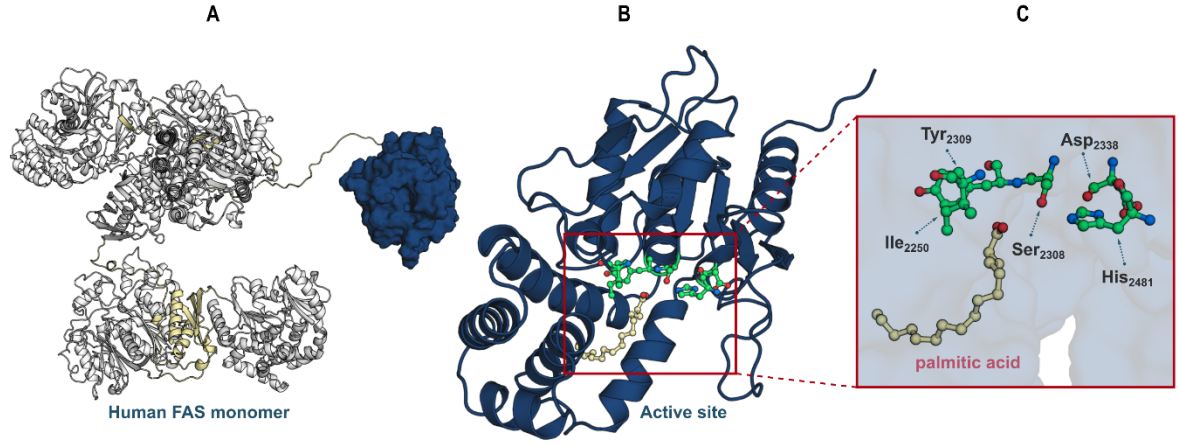


Figure 6| The TE domain of the animal (human) FAS [55]**.** A) Three-dimensional surface view of the TE domain within the human FAS monomer. B) TE secondary structure elements and active site region (within the red outlined box). The TE-substrate complex is taken from a computational model based on the crystallographic structure of the Homo sapiens TE domain (PDB ID: 2PX6) with a modelled palmitic acid molecule. C) Close-up of the active site cavity. The catalytic residues and the product are depicted in ball-and-stick representation.

Structure Overview

The TEs constitute a large superfamily of enzymes present in all living cells. Within its 25 families, and a recently proposed 26th one, [235] TEs catalyze the cleavage of thioester bonds on a wide array of cellular substrates: acyl-CoA, acyl-ACP, protein substrates, and acyl-glutathione. The TE of megasynthases, such as FAS and PKS, is characterized by a very conserved α/β hydrolase fold [231] with two subdomains – A and B.

The structure of the TE domain of animal FAS (Figure 6B) does not exhibit the canonical α/β hydrolase fold - its subdomain A displays only seven β strands, lacking the N-terminal β1 strand. As for the α portion of the canonical α/β hydrolase fold, TE's α portion is constituted by four α-helices flanking the half-barrel [53], with a loop insertion between strands β5 and β6. This loop insertion results in an atypical all-helical subdomain B near the catalytic triad site, unique to human FAS TE, and which has been suggested as crucial for TE's activity [236].

The catalytic triad of the α/β-hydrolase fold family involves a nucleophilic serine, a Asp/Glu and a histidine (Figure 6C). In the case of the animal TE catalytic triad, both the nucleophile (Ser2308) and the His2481 follow the canonical positioning, but the Asp2338 is located in the loop connecting subdomain B β5 strand to the subdomain A helix α5.

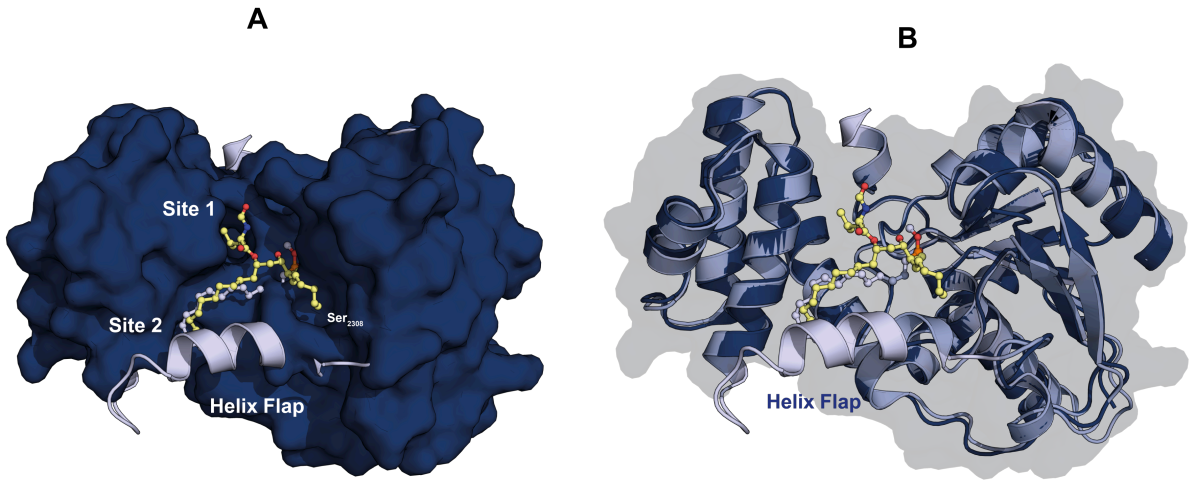


Figure 7| Structure of the binding site of the TE domain of the animal (human) FAS [55] A) Close-up three-dimensional surface view of the TE domain within the human FAS monomer (PDB ID: 2PX6). The active site cavity shows two distinct binding sites: Site 1, for short-length C chains, located at the interface between the two TE subdomains; and Site 2 for the C16-C18 substrates. The Helix Flap (PDB ID: 3TJM) highlighted in grey cartoon is responsible for the formation of the deep groove-tunnel Site 2 that harbors the C16-C18 substrate. The ligand Orlistat is represented in yellow ball-and-stick representation, with its peptidyl moiety protruding into the Site 1 resembling the behavior of short-length carbon chains. The C18 polyunsaturated fatty acid called methyl γ-linolenylfluorophosphonate (MGLFP, PDB ID: 3TJM) ligand is represented in grey ball-and-stick representation. For clarity, the active site residues are absent, except for the indication of the position of the Ser2308 that covalently binds the substrates. B) Cartoon representation of both the PDB ID: 2PX6 and 3TJM, aligned, displaying the helix flap in its disordered form (blue, PDB ID: 2PX6) and ordered form (grey, PDB ID: 3TJM).

Substrate Binding Site and Substrate Specificity

Animal FAS selectivity depends on TE's ability to interact with the C16 palmitoyl acyl substrate. The mechanism responsible for TE's substrate selectivity was first described by Chakravarty et al. in 2004: the TE mostly-hydrophobic substrate pocket was thought to be at the interface of the subdomains A and B, with about four to six carbons fitting into the interface cavity [53]. However, a resolved structure of the human

TE complex with its inhibitor Orlistat (PDB ID: 2PX6) challenged the previous proposal [236].

The structure (PDB ID: 2PX6) shows both the product state, as well as an acyl-enzyme intermediate. In the acyl-enzyme intermediate, it is possible to observe an ester linkage between Ser2308 and the C1 carbon atom of Orlistat, consistent with the nucleophilic attack at the carbonyl position. This frozen frame provided a detailed view into the catalytic mechanism of human FAS TE, and offered an unexpected insight into the acyl-enzyme complex. In the acyl-enzyme conformation, the electron density map shows Orlistat scaffold elements (peptidyl moiety, palmitic core, and hexanoyl tail) covering three different key regions of the TE domain in this step of the catalysis. However, this structure was missing an essential element for the substrate specificity mechanism: the amphipathic α-helix, called the flap (residues 2342-2355). This α-helix residue was first observed when Zhang and coworkers crystallized the human TE1 in complex with another substrate: a C18 polyunsaturated fatty acid called methyl γ-linolenylfluorophosphonate (MGLFP) (PDB ID: 3TJM) [237].

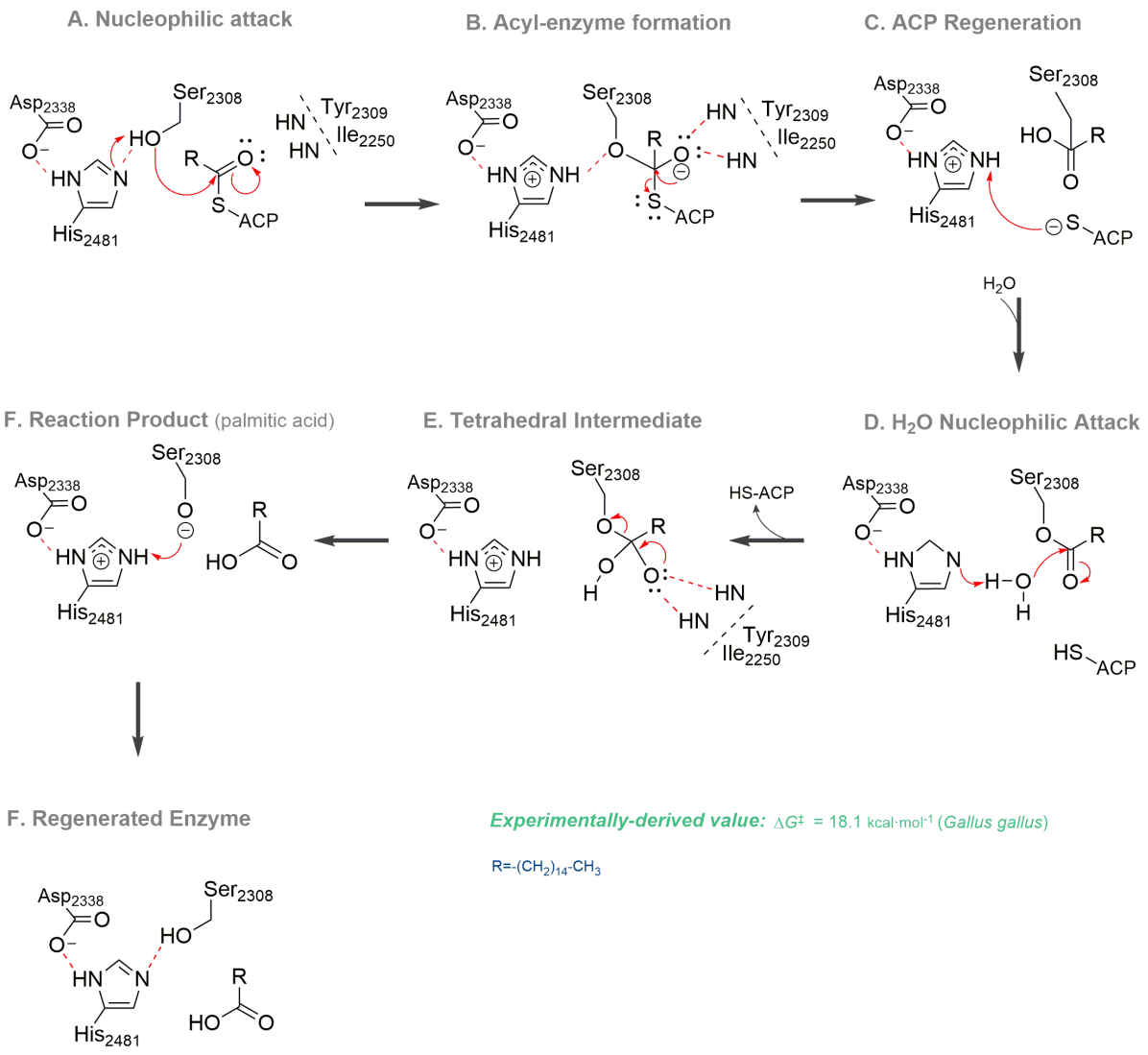
Comparison of the X-ray structures of human TE-Orlistat complex (PDB ID: 2PX6) [236] and human TE-MGLFP complex (PDB ID: 3TJM) reveals that, despite being very similar, there is a stark difference: the existence of an amphipathic α-helix, called the flap (residues 2342-2355) (Figure 7A and B), visible in the second crystal structure but previously unsolved in the first.

In sum, the works by Pemble et al. and Zhang et al. have shown the intricacies behind the mechanism for substrate specificity in TE. As PNS carries the substrate into the active site region, depending on its length, it can bind either in Site 1 at the interface between the two subdomains (up to 12 C atoms) or bind at Site 2 (longer than 12 C atoms) (Figure 7A). Spanning the two TE subdomains, Site 2 is thought to arise from the interplay between the disordered α-helix and the acyl-chain ligand, which causes concomitant ordering and involvement of a helix flap. The well-defined groove-tunnel Site 2 starts near the Ser2308 in the main subdomain, and terminates deep in the tunnel in the small subdomain formed by the triangulation of the three helices - the

helix flap, α5 and α8. This finding substantiates the importance of the two subdomains combined. Moreover, at the end of Site 2, the residues Arg2428, Glu2431 and Tyr2424 (helix α8) are thought to be responsible for docking the C16-C18 substrate into place while generating a barrier to longer substrates.

The Catalytic Mechanism

For TE's effective catalytic action and release of 16 or 18 carbon long products, three key moments need to take place: *i)* transient protein:protein interactions between ACP and TE need to be established; *ii)* recognition of the substrate chain length; *iii)* cleavage of the fatty acid chain from the ACP, releasing the product into the cytosol. TE's active site exhibits a canonical catalytic triad: Ser2308, His2841, Asp2338 [236], similar to the canonical mechanism of hydrolases with an identical catalytic triad [53,238], the hydrolysis of the substrate from the ACP proceeds in two distinct steps.



Scheme 1| Catalytic mechanism proposed for FAS TE domain. Proposed mechanistic pathways for the TE domain catalytic mechanism. ("R" stands for the acyl moiety of C16 that will be cleaved by the TE domain. The residues are numbered according to the human TE domain). The Gibbs activation energy of 18.1 kcal·mol^{-1} was derived from the k_{cat} value of 0.32 s^{-1} (25 °C) [239] using the Transition State Theory.

Firstly, the catalytic Ser2308 hydroxyl should be deprotonated by His2481, resulting in an enhanced nucleophilic character at the Oγ of the Ser. The nucleophilic Oγ then attacks the carbonyl carbon of the substrate, forming an acyl-enzyme tetrahedral intermediate, with the resulting acyl oxyanion stabilized by an oxyanion hole formed by the backbone amide nitrogens of Tyr2309 and Ile2250. The resulting nucleophilic PNS can then deprotonate the positively charged His2481, which is then available to deprotonate a water molecule to form a hydroxide nucleophile prone to attack the

carbonyl carbon at the acyl-enzyme ester, resulting in a second negatively charged tetrahedral intermediate. Finally, the unstable intermediate is resolved upon deprotonation of His2481 by Ser2308, regenerating the active site and releasing the carboxylic acid product (Scheme 1).

3.3.3. Methods

Acyl-Enzyme Model Preparation

The computational model in the present study was based on the X-ray structure of human TE in complex with the FDA-approved Orlistat inhibitor (PDB ID: 2PX6, 2.30Å). This structure, as previously explained, exhibits Orlistat in each of the active sites of two TE monomers – in one monomer the Orlistat forms a stable acyl-enzyme, and in the second it can be found as its hydrolyzed product. Orlistat's structure is similar to palmitate, and binds almost exclusively to a hydrophobic surface channel formed by the subdomain B, as it would be expected of a mostly hydrophobic palmitate molecule.

In the PDB ID: 2PX6 crystal structure of the human TE-Orlistat complex, three loop regions had poor electron density and were absent in the original X-ray structure: loop I (residues 2326–2328, missing in chain A only); loop II (residues 2344–2360) bridges subdomain A and subdomain B; and loop III (residues 2450–2460) is near the catalytic triad linking β6 to β7. Due to the proximity of these loop regions to the active site, it was crucial to model these regions through homology modeling.

Using SWISS-Model for homology modeling, it was possible to model the missing loop regions in the human TE-Orlistat complex using the crystal structure of the human TE-MGLFP complex (PDB ID: 3TJM, 1.48Å). Despite of the improved resolution of the human TE-MGLFP complex, it did not provide the two distinct states of human TE as seen in the human TE-Orlistat structure. These intermediate and product stages of the reaction should prove useful for QM/MM studies. Thus, we opted to use human TE-Orlistat as the scaffold of our model, and used human TE-MGLFP complex as a template to model the absent loop regions. One of the three loop regions, loop II,

corresponds to the helix flap highlighted in Figure 7. This flap helix (residues 2342-2355) establishes several complementary interactions with the γ-linolenyl chain of the MGLFP substrate, creating a hydrophobic groove-tunnel that harnesses the substrate; interestingly, the flap helix is absent in the ligand-free structure determined by Zhang et al., suggesting that the formation of the gatekeeper α-helix result from a disordered polypeptide segment upon binding of the γ-linolenyl moiety of the substrate (MGLFP) [237].

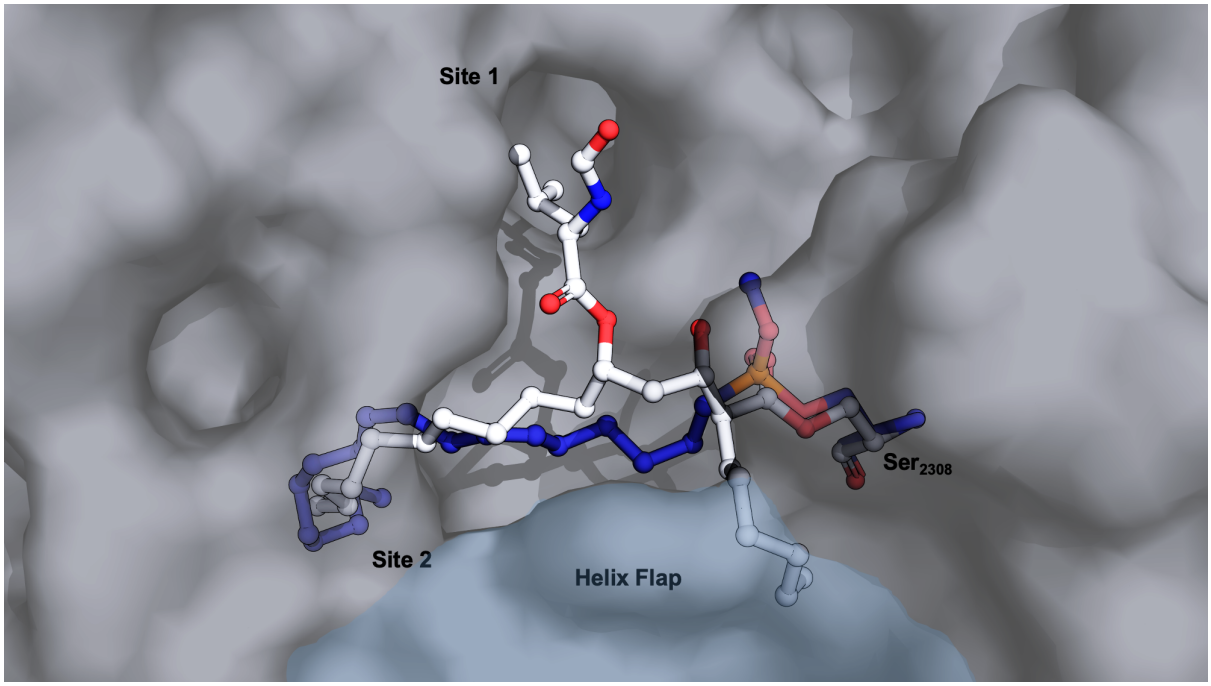
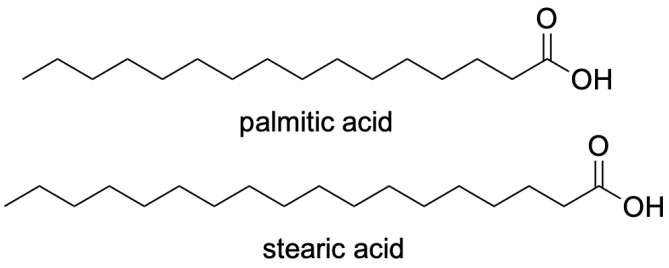


Figure 8| Representation of the TE domain's active site, with MGLFP and Orlistat ligands bound to Ser2308.

This work considered the protonation states for all amino acids at pH 7.0, according to PROPKA 3.1 predictions [196,240]. In particular, in the acyl-enzyme state, the His2481 catalytic residue was modelled in the neutral form protonated at Nδ, while the Asp2338 was modelled in its negatively charged form, agreeing with its solution pK_a. These protonation states follow those proposed for other α/β hydrolases with similar Ser-His-Asp/Glu triads [235], where His2481 is proposed to deprotonate the nucleophilic water, and the resulting imidazolium is stabilized by Asp2338.



Scheme 2| TE domain products: palmitic acid (C16) and stearic acid (C18).

The natural substrate of human FAS TE is palmitoyl (C16) or stearate (C18) bound to the PNS cofactor of ACP. The target of TE is the thioester bond found between the palmitoyl moiety and the PNS arm of ACP. At the acyl-enzyme state, PNS is no longer present at the active site, and the palmitoyl moiety is bound to Ser2308. In the employed X-ray structure [241], the acyl-enzyme complex displays the inhibitor Orlistat present at the active site. The Orlistat molecule pose at the active site of acyl-TE served as a template for the correct positioning of the palmitoyl. Modelling of the Serine-Palmitoyl (SPL) modified residue was performed with GaussView5 [242].
The parameters for the SPL modified residue were determined from an optimized structure at the HF/6-31G(d) level of theory. These were obtained with the Antechamber tool, as implemented in AMBER12 [179], using the GAFF force field [243] to derive the intramolecular and Lennard-Jones parameters and the RESP (Restrained Electrostatic Potential) [244] method to derive atomic charges at the HF/6-31G(d) level of theory. Modelling of the residue SPL was subsequently validated by MD simulations to ensure the quality of the resulting structure. As the employed modeled structure is in a quasi-reactive conformation from the crystal structure, physical constraints were applied to the active site residues (Ser2308 bound to the palmitoyl moiety, His2481, Asp2338, Tyr2309 and Ile2250). The modelling protocol adopted in the study ensures the quality of the resulting structure.

The enzyme was parameterized using the AMBER FF99SB force field [199,245], while ions and hydrogen atoms were added using the XLeap module of Amber12 software [179]. Finally, TIP3P water molecules [181] were added within 16.0 Å from the modelled system to fill a rectangular box.

Molecular Dynamics

cMD calculations were performed using the AMBER 12 software [179], after an energy minimization protocol of *i)* water molecules, followed by *ii)* covalent hydrogen bonds, *iii)* sidechains, *iv)* loops, *v)* active site, *vi)* serine-palmitoyl modified residue, and finally, *vii)* the entire system. The minimization steps were applied for model optimization and to avoid tensions and clashes derived from the previous modelling. To constrain bonds involving hydrogen atoms in all cMD calculations, the SHAKE algorithm [246] was employed, and an integration step of 2 fs was used. Annealing was performed from 0 to 310 K during 100 ps at constant volume, followed by another 100 ps at 310 K in the *NVT* ensemble. During this phase, the protein backbone remained restrained. The above-mentioned restraints employed harmonic potential force constants of 50 kcal·mol^{-1} Å^{-2}. cMD productions were then performed to analyze the conformations of the active site region in an *NPT* ensemble at 1 bar and 310 K. The Berendsen barostat and the Langevin thermostat were used throughout the simulations.

For the production phase of the cMD, the protocol was: *i)* 14 ns with protein backbone and substrate restrained; *ii)* 14 ns with protein backbone restrained, except for the active site residues Ser2308 bound to the palmitoyl moiety, His2481, Asp2338, Tyr2309 and Ile2250); *iii)* followed by 40 ns of unrestrained cMD.

3.3.4. Results

For the choice of the MD production frame from which the QM/MM model of the intermediate TE-Palmitoyl complex would be built, we opted for the last frame of the 14 ns production phase with protein backbone constrained, except for the active site residues Ser2308 bound to the palmitoyl moiety, His2481, Asp2338, Tyr2309 and Ile2250. As the X-ray structure was in a quasi-reactive conformation, by keeping the protein backbone restrained and the active site free of constraints, we are able to maintain the conformation while allowing the active site residues and modeled substrate to relax and prevent any steric hindrances.

The proposal for the deacylation and regeneration of the TE domain requires the presence of a water molecule at the active site. The activated water is responsible for the nucleophilic attack at the ester carbon of the enzyme-PLM adduct. We have evaluated the presence of water molecules at the active site, using the criterium: the water molecule must be present within a 3 Å radius from both the Nε of His2481 and oriented towards the carbon of the ester of the palmitoyl.

The QM layer of the TE-Palmitoyl complex model was defined based on the active site residues Ser2308, His2481, Asp2338, Tyr2309 and Ile2250, and it includes a water molecule that fits the criterium. We also decided to include two other water molecules that form hydrogen bonds to the nucleophilic water (Figure 9).

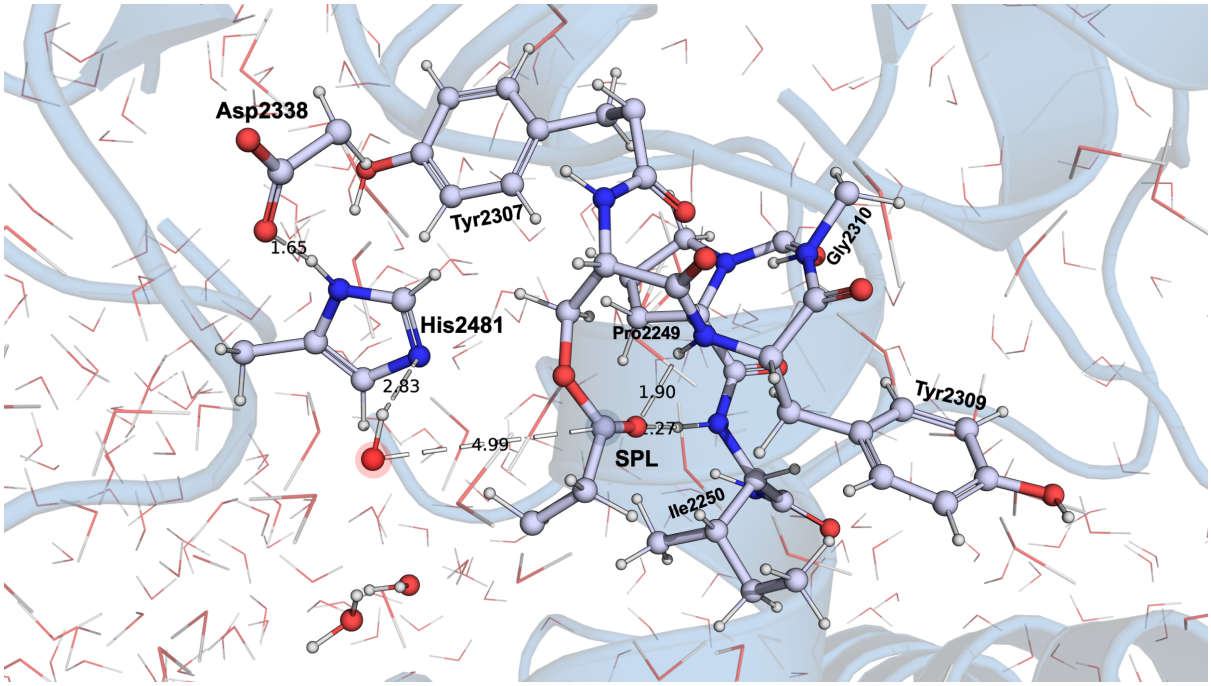


Figure 9| QM/MM model of the TE-Palmitoyl intermediate. Transparent cartoon representation of the TE protein, as well as the transparent stick representation of water molecules, depict the considered MM layer. The atoms represented in ball and stick belong to the QM layer. The QM layer includes the catalytic triad Ser2308, His2481, Asp2338 and the palmitoyl moiety, alongside Tyr2309, Ile2250 that stabilize the oxyanion hole. Tyr2307, Pro2249 and Gly2310 establish a network of hydrogen bonds that contribute to the stabilization and rigidity of the active site and therefore should be included in the QM layer.

3.3.5. Future steps

With the proposed QM/MM model, we hope to initiate a QM/MM study of the catalytic mechanism of the TE domain of hFAS. By exploring the reaction coordinate of the nucleophilic attack by the activated water at the Intermediate sate of the TE-Palmitoyl complex, we aim to characterize its transition state. By doing so, it will be possible to carry out an IRC calculation from the transition state to identify its corresponding minima (the acyl-enzyme intermediate and the palmitate product). By understanding the catalytic mechanism of the TE domain of the human FAS, it will be possible to fully unveil the potential of TE's specificity and hydrolase activity for production of fatty acid derivatives of specific size.

Since members of the α/β hydrolases superfamily are able to catalyze the hydrolysis and cyclization of a wide variety of substrates, it would also be important to understand their selectivity towards specific molecules. In that regard, for the human FAS TE, the helix flap found upon substrate binding may act as a controlling device that provides a hydrophobic surface channel that effectively binds the right-sized substrate.

To explore the behavior of the helix flap and its interaction with different-sized substrates, the binding of ACP to the TE domain of FAS, throughout the many stages of the elongation cycle, could be assessed through the combination of protein-protein docking, molecular modeling and extensive MD simulations. Some possible attempts are summarized below.

ACP domain without substrate	*ACP domain + PNS with a C6 substrate*	*ACP domain + PNS with a C16 substrate*
Docking of ACP onto TE using HADDOCK	Docking of ACP onto TE using HADDOCK	Docking of ACP onto TE using HADDOCK
Assembly of ACP:TE complex	Assembly of ACP:(TE-C6 substrate) complex	Assembly of ACP:(TE-C16 substrate) complex
MD simulation up to μs	MD simulation up to μs	MD simulation up to μs
Five Replicas	Five Replicas	Five Replicas
Clustering	Clustering	Clustering
Identification of predominant poses and comparison of helix flap conformations		

Table 1| Experimental proposal for the study of the behavior of the helix flap and its interaction with different-sized substrates, the binding of ACP to the TE domain of FAS, throughout the many stages of the elongation cycle.

3.4. Outcomes and Future Prospects

The study of the FAS system done over the course of this thesis established a solid framework. Its study began with the 1.5 million atoms model of hFAS, encompassing its 14 catalytic domains in two monomers. Due to the complexity and mobility of this megaenzyme, it is a challenging protein to obtain an X-ray crystallographic structure of, delaying possible scientific breakthroughs. The *in silico* approaches - homology modeling and MD simulations - employed in our project allowed for a refined and plausible structure of hFAS. We were also able to determine the docking poses of ACP and five of the catalytic domains of hFAS (MAT, KS, KR, DH and ER), providing insight into the residues that are involved in domain-domain contacts, as well as into the active site residues when the ACP is docked at each of the domains. These checkpoints have allowed for the initiation of the catalytic mechanism studies of domains MAT, KS, KR, DH and ER performed by our group.

In the final stages of this thesis, we have further extended our studies of hFAS with the proposal of a model for upcoming QM/MM studies of TE's catalytic mechanism.

With the growing body of literature available on the FAS system, we have delved into this system in depth and authored a comprehensive review named "Animal Fatty Acid Synthase: a Chemical Nanofactory" published in 2020 on Chemical Reviews. Due to its length, the first page of the review is available in subchapter 3.4.1, and the integral version of the review is available in the Appendix section.

Due to the shared evolutionary history of FAS and PKS systems, a review to summarize the state-of-the-art knowledge on the PKS systems was also published, and it is available in subchapter 3.4.2.

We hope that our survey allows established a solid ground for future research on these exquisite systems, whose potential extends from pharmaceuticals to biofuels.

3.4.1. Animal Fatty Acid Synthase: a Chemical Nanofactory

Rui P. P. Neves[a], Pedro Ferreira[a], Fabiola E. Medina [b], Pedro Paiva[a], João P. M. Sousa[a], Matilde F. Viegas[a], Maria J. Ramos[a], and Pedro A. Fernandes[a]

[a] LAQV@REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

[b] Departamento de Ciencias Químicas, Facultad de Ciencias Exactas, Universidad Andres Bello, Autopista Concepción-Talcahuano, 7100 Talcahuano, Chile

---

Paiva, P., Medina, F. E., **Viegas, M. F.**, Ferreira, P., Neves, R. P. P., Sousa, J. P. M., Ramos, M. J., & Fernandes, P. A. (2021). ***Animal fatty acid synthase: A chemical nanofactory***. In Chemical Reviews, 121(15), 9502–9553. https://doi.org/10.1021/acs.chemrev.1c00147

3.4.1.1. Abstract

Fatty acids are crucial molecules for most living beings, very well spread and conserved across species. These molecules play a role in energy storage, cell membrane architecture, and cell signaling, the latter through their derivative metabolites. *De novo* synthesis of fatty acids is a complex chemical process that can be achieved either by a metabolic pathway built by a sequence of individual enzymes, such as in most bacteria, or by a single, large multi-enzyme, which incorporates all the chemical capabilities of the metabolic pathway, such as in animals and fungi, and in some bacteria. Here we focus on the multienzymes, specifically in the animal fatty acid synthase (FAS). We start by providing a historical overview of this vast field of research. We follow by describing the extraordinary architecture of animal FAS, a homodimeric multi-enzyme with seven different active sites per dimer, including a carrier protein that carries the intermediates from one active site to the next. We then delve into this multi-enzyme's detailed chemistry and critically discuss the current knowledge on the chemical mechanism of each of the steps necessary to synthesize a single fatty acid molecule with atomic detail. In line with this, we discuss the potential and achieved FAS applications in biotechnology, as biosynthetic machines, and compare them with their homologous polyketide synthases, which are also finding wide applications in the same field. Finally, we discuss some open questions on the architecture of FAS, such as their peculiar substrate-shuttling arm, and describe possible reasons for the emergence of large megasynthases during evolution, questions that have fascinated biochemists from long ago but are still far from answered and understood.

Keywords: Fatty Acid Synthesis, Fatty Acid Synthase, Enzymatic Catalysis, Megaenzyme.

CHEMICAL REVIEWS
pubs.acs.org/CR
Review

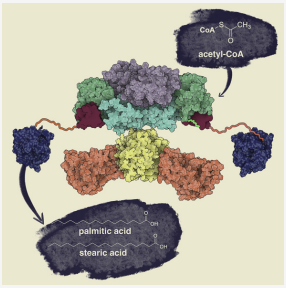
Animal Fatty Acid Synthase: A Chemical Nanofactory

Pedro Paiva,[⊥] Fabiola E. Medina,[⊥] Matilde Viegas,[⊥] Pedro Ferreira,[⊥] Rui P. P. Neves,[⊥] João P. M. Sousa,[⊥] Maria J. Ramos, and Pedro A. Fernandes*

Cite This: https://doi.org/10.1021/acs.chemrev.1c00147
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Fatty acids are crucial molecules for most living beings, very well spread and conserved across species. These molecules play a role in energy storage, cell membrane architecture, and cell signaling, the latter through their derivative metabolites. *De novo* synthesis of fatty acids is a complex chemical process that can be achieved either by a metabolic pathway built by a sequence of individual enzymes, such as in most bacteria, or by a single, large multi-enzyme, which incorporates all the chemical capabilities of the metabolic pathway, such as in animals and fungi, and in some bacteria. Here we focus on the multi-enzymes, specifically in the animal fatty acid synthase (FAS). We start by providing a historical overview of this vast field of research. We follow by describing the extraordinary architecture of animal FAS, a homodimeric multi-enzyme with seven different active sites per dimer, including a carrier protein that carries the intermediates from one active site to the next. We then delve into this multi-enzyme's detailed chemistry and critically discuss the current knowledge on the chemical mechanism of each of the steps necessary to synthesize a single fatty acid molecule with atomic detail. In line with this, we discuss the potential and achieved FAS applications in biotechnology, as biosynthetic machines, and compare them with their homologous polyketide synthases, which are also finding wide applications in the same field. Finally, we discuss some open questions on the architecture of FAS, such as their peculiar substrate-shuttling arm, and describe possible reasons for the emergence of large megasynthases during evolution, questions that have fascinated biochemists from long ago but are still far from answered and understood.



Downloaded via Pedro Paiva on June 23, 2021 at 22:23:26 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

CONTENTS

1. Historical Perspective B
1.1. Fatty Acid Biosynthesis B
1.2. Fatty Acid Synthase, Cancer, and Obesity C
2. Animal Fatty Acid Synthase: Architecture and Function D
2.1. Architecture of Animal FAS D
2.1.1. Expression, Folding, and Assembly Elasticity D
2.1.2. Overall Architecture and Catalysis D
2.1.3. Head-to-Tail vs Head-to-Head Orientation E
2.1.4. Structural Scaffolding and Flexibility E
2.2. Physiological Role of Animal FAS and Its Role in Therapeutics F
2.2.1. Mechanisms of Animal FAS Expression and Implications in Disease F
2.2.2. Drug Design Targeting Animal FAS G
3. The Catalytic Mechanism of Animal FAS H
3.1. Reaction Cycle of the MAT Domain H
3.1.1. Substrate Binding and Promiscuity H
3.1.2. Active-Site Structure and Catalytic Mechanism I
3.2. Reaction Cycle of the KS Domain K
3.2.1. Structure and Function of the Animal KS Domain K
3.2.2. Catalytic Machinery of the KS Domain L
3.2.3. Role of the Animal KS Domain in the Fatty Acid Chain-Length Regulation P
3.3. Reaction Cycle of the KR Domain Q
3.3.1. Biochemical Characterization of Animal KR Q
3.3.2. Active-Site Structure and Catalytic Mechanism of KR Q
3.4. Reaction Cycle of the DH Domain S
3.4.1. Biochemical and Structural Characterization S
3.4.2. Active-Site Structure and Catalytic Mechanism S
3.5. Reaction Cycle of the ER Domain T
3.5.1. Biochemical and Structural Characterization T
3.5.2. Active-Site Structure and Catalytic Mechanism T
3.6. Reaction Cycle of the TE Domain W

Received: February 19, 2021

© XXXX American Chemical Society

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

3.4.2. Engineering of PKS megaenzymes: a promising way to biosynthesize high-value active molecules

Rui P. P. Neves[a], Pedro Ferreira[a], Fabiola E. Medina [b], Pedro Paiva[a], João P. M. Sousa[a], Matilde F. Viegas[a], Maria J. Ramos[a], and Pedro A. Fernandes[a]

[a] LAQV@REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

[b] Departamento de Ciencias Químicas, Facultad de Ciencias Exactas, Universidad Andres Bello, Autopista Concepción-Talcahuano, 7100 Talcahuano, Chile

---

Neves, R. P. P., Ferreira, P., Medina, F. E., Paiva, P., Sousa, J. P. M., **Viegas, M. F.**, Fernandes, P. A., & Ramos, M. J. (2021). ***Engineering of PKS megaenzymes—A promising way to biosynthesize high-value active molecules***. In Topics in Catalysis. https://doi.org/10.1007/s11244-021-01490-5

3.4.2.1. Abstract

Polyketide synthases (PKS) produce polyketides, a large group of secondary metabolites produced in bacteria, fungi, plants and certain animal families. PKSs have a huge potential to biosynthesize high-value active pharmaceuticals, agrochemicals, biofuels and industrial chemicals, among others. Additionally, they present ideal opportunities to engineer new polyketides, and the latter chemical novelty can result in even better, or at least alternative, active molecules to the existing ones. We present the structure, catalysis and altogether machinery of the PKS, as well as their enzyme engineering and applications in biotechnology. We also briefly discuss the current progress on the study of PKSs by computational chemists, and the fact that PKS reactivity and catalysis remain, presently, almost open fields for the computational chemist to explore.

Keywords: Polyketide Synthases, Enzyme Engineering, Biotechnology, Polyketide Synthesis.

Topics in Catalysis
https://doi.org/10.1007/s11244-021-01490-5

ORIGINAL PAPER

Engineering of PKS Megaenzymes—A Promising Way to Biosynthesize High-Value Active Molecules

Rui P. P. Neves[1] · Pedro Ferreira[1] · Fabiola E. Medina[2] · Pedro Paiva[1] · João P. M. Sousa[1] · Matilde F. Viegas[1] · Pedro A. Fernandes[1] · Maria J. Ramos[1]

Accepted: 29 July 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
Polyketide synthases (PKS) produce polyketides, a large group of secondary metabolites produced in bacteria, fungi, plants and certain animal families. PKSs have a huge potential to biosynthesize high-value active pharmaceuticals, agrochemicals, biofuels and industrial chemicals, among others. Additionally, they present ideal opportunities to engineer new polyketides, and the latter chemical novelty can result in even better, or at least alternative, active molecules to the existing ones. We present the structure, catalysis and altogether machinery of the PKS, as well as their enzyme engineering and applications in biotechnology. We also briefly discuss the current progress on the study of PKSs by computational chemists, and the fact that PKS reactivity and catalysis remain, presently, almost open fields for the computational chemist to explore.

1 The PKS Enzymes—Structure and Catalysis

Polyketide synthases (PKS) are enzymes that produce the carbon scaffolds for molecules, *i.e.* they are a family of enzymes involved in the secondary metabolism that catalyze condensation of acetyl- and malonyl-CoA primary metabolites to synthesize β-ketoacyl polymers. PKSs can be found as discrete monofunctional enzymes (type-II PKSs) and as multifunctional megasynthases (type-I PKSs) amounting to several MDa in weight, depending on the organism of expression and the output product (polyketides) [1–4].

Given that PKS and FAS (fatty acid synthase) systems are evolutionarily related, they are very often compared in both their architecture and catalysis. However, PKS have actually evolved in a more sophisticated way, as they are capable of catalyzing the synthesis of a large diversity of polyketides, with varying carbon reduction states and degree of aromaticity [5–10]. Such polyketide diversity is achieved due to the variation of carbon processing enzymes (type-II) or domains (type-I) ketoreductases (KR), dehydratases (DH), enoyl reductases (ER) and methyl transferase (ME), while retaining a canonical condensation step catalyzed by acyl-CoA transferases (AT) and ketosynthases (KS) [1, 11]. In tandem, it allows for the ketide extender groups to be converted to alcohols, alkenes or alkanes [2–4]. Furthermore, some PKSs can also accept varied acyl-CoA starters [12, 13] or include additional catalytic domains/enzymes, e.g., aromatases (AR), which broadens the variety of synthesized polyketide secondary metabolites. Finally, PKSs can also function iteratively, encompassing cycles of polyketide condensation and carbon processing, or in modular assemblies, where each step of polyketide elongation is carried out by successive PKS megasynthase modules bonded to one another covalently or non-covalently. Despite that there are numerous examples of structural data from PKS enzymes (type-II) in the Protein Data Bank, less than a dozen examples can be found for PKS megaenzymes (type-I). Nevertheless, this data has allowed significant advances in the field up to now [14–17].

PKS biosynthesis is a complex process in which each PKS assembles an individual product dependent on the selection of acyl-CoA units and their level of reduction, thus ensuing shaping of the final polyketide. The process is carried out iteratively and broadly follows three steps—starting, elongation and termination (refer to Fig. 1 for an example of a type-II PKS involved in actinorhodin biosynthesis)

✉ Maria J. Ramos
mjramos@fc.up.pt

[1] LAQV, REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal

[2] Departamento de Ciencias Químicas, Facultad de Ciencias Exactas, Universidad Andres Bello, Autopista Concepción-Talcahuano, 7100 Talcahuano, Chile

Published online: 04 September 2021

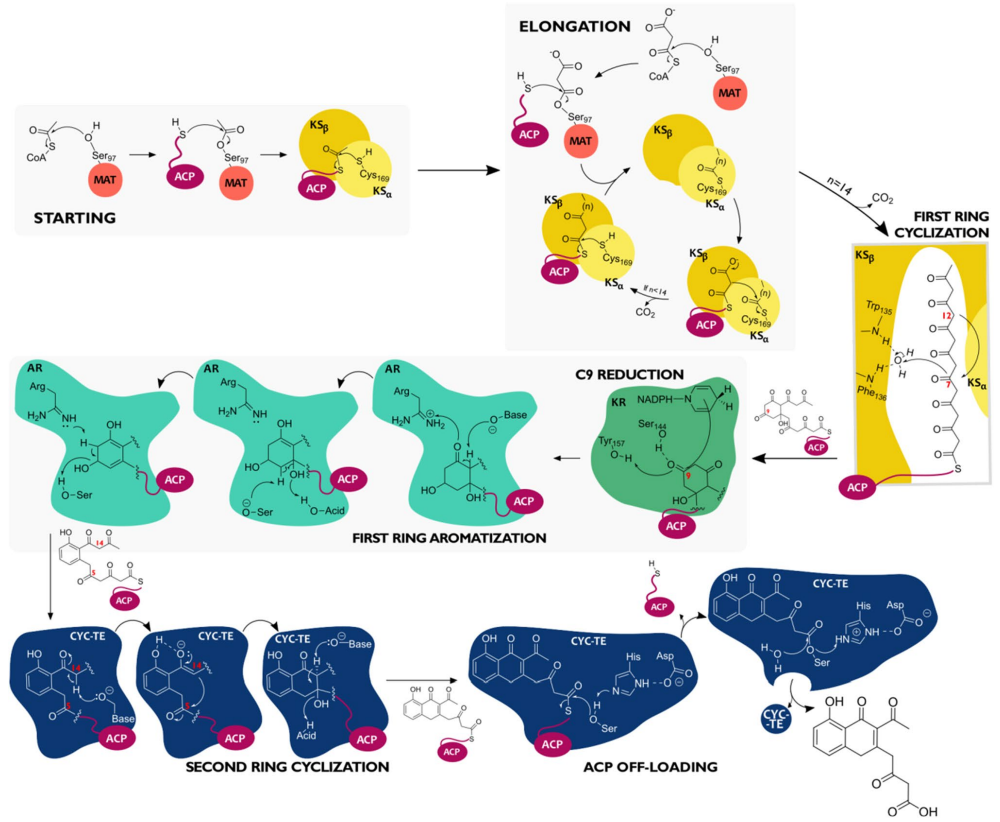


Fig. 1 The steps of actinorhodin biosynthesis carried out by type-II PKSs [18–23]. The starting step requires acetyl-CoA that is loaded into a malonyl/acetyl-CoA transferase (MAT) and then transferred to the ketosynthase (KS_α:KS_β) by the acyl-carrier protein (ACP). The elongation step comprises the loading of malonyl-CoA extenders into MAT and their loading into KS_α:KS_β by ACP until a 16-carbon polyketide is formed. After a first cyclization involving the C7-ketone and C12, the C9-ketone is reduced to C9-hydroxyl by a ketoreductase (KR)—C9 reduction–, and an aromatization reaction is catalyzed by an aromatase (AR)—first ring aromatization. A second ring cyclization is catalyzed by a cyclase-thioesterase domain (CYC-TE) between the C5 and C14 of the polyketide, and the ACP off-loading is finally catalyzed through a hydrolysis by the CYC-TE. The final actinorhodin is synthesized by non-PKS enzymes. Mechanisms based on canonical proposals (such as those by AR and CYC-TE) do not include residue numbers

[18–20]. Along these three steps, transacylations mediate the growth of the polyketide molecule, until it is finally released by hydrolysis or cyclization.

Given the structural resemblance and the evolutionary relatedness of FAS and PKS, the catalytic mechanisms of PKS are expected to be quite similar to those of FAS. In recent years, we have been using hybrid quantum mechanics/molecular mechanics (QM/MM) calculations to unveil the reaction mechanisms of the catalytic domains of animal FAS, namely the malonyl-acetyl transferase (MAT) loading domain [24, 25], and the carbon processing domains KR, DH and ER [26–28]. In line with available structural data, the catalytic residues validated to date in FAS are conserved across every PKS and FAS enzymes, supporting the hypothesis that the chemical steps of polyketide and fatty acid syntheses might occur similarly. These systematic studies suggested that the rate-limiting step of the reaction catalyzed by FAS is likely to reside in large domain rearrangements to ensure substrate transport, and not in the chemical reactions occurring at each domain, whose activation Gibbs energies fall within 15.6–18.8 kcal·mol^{-1}.[29] Hence, we believe that these

results may be transferable for PKS megasynthases. Figures 1, 2 summarize polyketide biosynthetic pathways for type-II and type-I PKSs, respectively.

At the starting step, a specific acyl-CoA group is recruited by the AT domain, and the resulting unit constitutes an initial substrate. Based in current knowledge on fatty acid synthesis by animal FAS, polyketide synthesis initiates with the loading of a covalent ACP-phosphopantetheine (PNS) complex by the AT domain, enclosing a conserved Ser-His dyad responsible for both loading and unloading of the AT domain. The loading of AT by an acyl-CoA occurs through a rate-limiting step nucleophilic attack of the conserved Ser of the AT to the thioester bridging the CoA and the acyl unit, upon an asynchronous deprotonation of the Ser by the His of the dyad [24]. The resulting covalent AT-acyl complex is then deacylated (i.e., unloaded) by the ACP-PNS complex after a second nucleophilic attack by the sulfur of PNS [25]. Polyspecificity of the starter units, *i.e.* the promiscuity of the AT domain to recruit multiple different starter groups, is more commonly observed at this stage than at later stages, which renders the starting of biosynthesis a promising step for the generation of novel polyketides [32]. At the elongation step, the elongation group, usually malonyl- or methylmalonyl-CoA, is transported from the AT domain onto the KS domain by the acyl carrier protein (ACP) shuttling domain. The polyspecificity of the elongation group is limited and its prediction is hardly straightforward. The ACP-bound elongation group reacts in a Claisen condensation reaction with the KS-bound polyketide chain and, as a result, the KS domain becomes free to participate in another reaction. The diversity of KS enzymes is impressive in the PKS family of enzymes; however, KS of type-I PKSs conserve a Cys-His-His triad similar to that of animal FAS [33, 34]. Despite that the reaction mechanism taking place at KS remains elusive, the Cys of the conserved triad is proposed to catalyze the polyketide transfer from ACP-PNS to KS, prior to the cargo of ACP at AT, and the two His are proposed to stabilize the carbanion formed upon the Claisen condensation reaction through an oxyanion hole mechanism [35, 36]. Subsequently, the ACP gets bound to the elongated polyketide chain and becomes ready to repeat another elongation cycle, until the polyketide chain is completed and the termination step is reached. However, before it takes place, the polyketide chain can be altered also by the other domains—e.g., the KR domain may reduce the β-keto group to a β-hydroxy group [26], the DH domain may dehydrate the polyketide chain into an α-β-unsaturated alkene [27], the ER domain may reduce the α-β-double-bond of the β-keto-acyl chain into a single-bond [28], or the ME domain may methylate the β-carbon of the growing polyketide chain (refer to Fig. 2 for an example of the synthesis of the dihydromonacolin L acid from the lovastatin biosynthetic pathway by the iterative type-I PKS LovBC and LovG) [17, 30, 31]. Therefore, diversity into the polyketide scaffold is introduced in many ways by the PKSs, from the manipulation of the starter and extender modules to the action of different domains, all collaborating to introduce an enlarged chemistry into the polyketide backbone. Eventually, at the termination step, the complete polyketide chain is released from the ACP domain (*i.e.*, hydrolyzed) by the thioesterase (TE) domain.

The comprehension of the biosynthetic pathway catalyzed by PKSs has been recently made stronger by studies that have shown broadly how the PKS assembly line works for the particular class of modular PKSs. In fact, recent in vitro kinetic measurements of ^{14}C-labeled polyketides from different assemblies of modules from the 6-DEBS synthase (6-deoxyerythronolide B synthase), a prototypical modular PKS, indicated that the Claisen condensation by each KS domain acts as a turnstile and contributes to the vectorial nature of modular PKSs [34, 37]. Furthermore, structural characterization of *C*- and *N*-terminal docking domains in modules from DEBS and virginiamycin polyketide synthase pointed towards specific protein:protein interactions between these, which seem to contribute for intermodular polyketide transfer and module dimerization (Fig. 3, on the left) [15, 38]. The role of the latter in the vectorial nature of modular PKSs was again stressed from cryo-electron microscopy studies of the module 5 of pikromycin PKS (Fig. 3, on the right) [14], where dimerization was referred to contribute for the selectivity of ACP towards the KS in each module.

In sum, the turnstile mechanism controls chain growth by preventing the acylation of the KS domain of each module by an incipient polyketide chain, before the product of the earlier catalytic cycle has been passed to the ACP domain. However, despite these reported advances, the overall dynamics of modular PKSs remains hampered by the lack of structural detail of the whole modular PKS. The dynamics of modular PKS is currently thought to be the result of a complex synergy between protein:protein recognition among catalytic domains including the ACP domain, resulting in a very broad variety of polyketide products as previously referred [39, 40].

2 The PKS Machinery (Megasynthases vs. Monofunctional Synthases)

PKS systems can be either megasynthases (type-I systems) or monofunctional enzymes (type-II systems) [5]. Moreover, a particular class of PKS systems (type-III) has also been recognized, involving a family of homodimeric KS that are able to carry out condensation reactions of polyketide

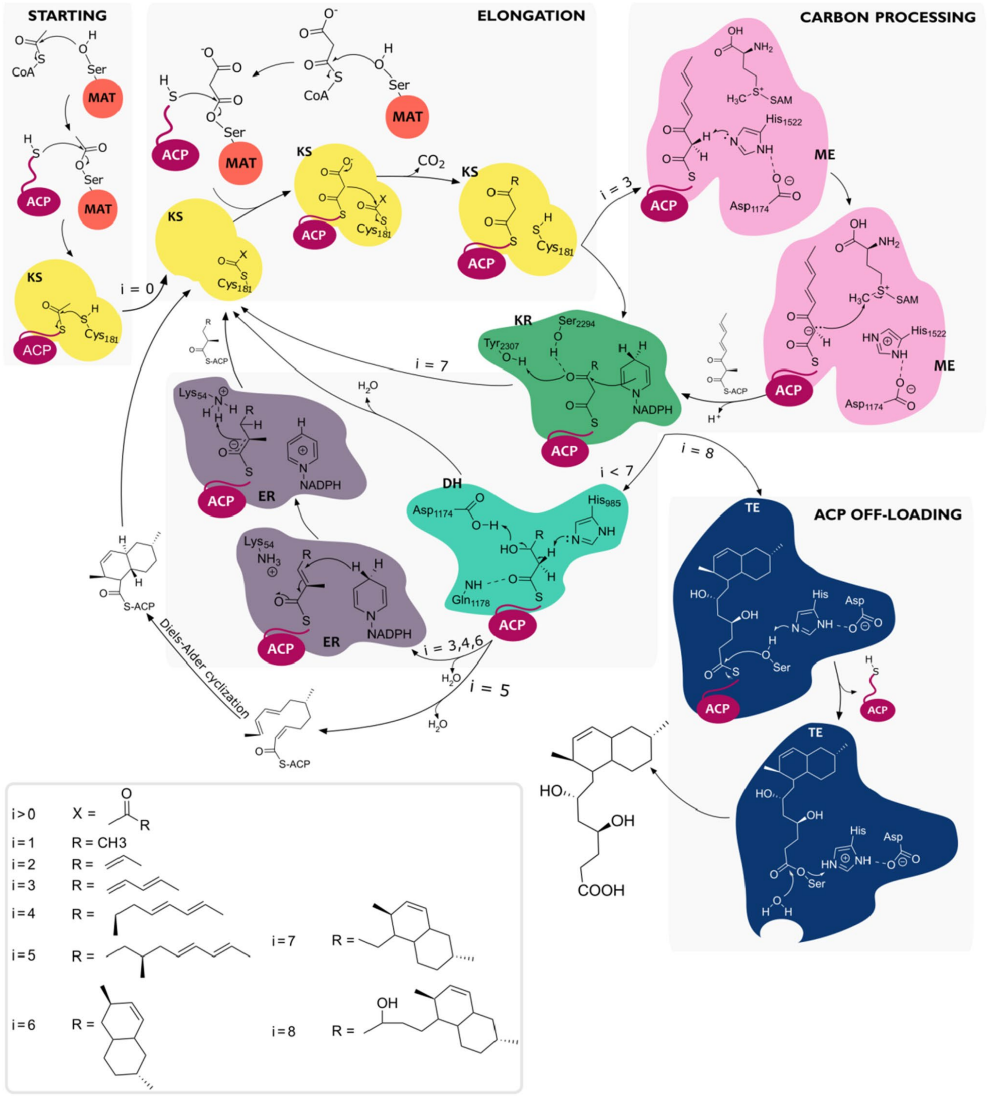


Fig. 2 The steps of lovastatin biosynthesis carried out by the type-I PKS LovBC [17, 30, 31]. The starting step requires acetyl-CoA that is loaded into a malonyl/acetyl-CoA transferase (MAT) and then transferred to the ketosynthase (KS) by the acyl-carrier protein (ACP). The elongation step comprises the loading of malonyl-CoA extenders into MAT and their loading into KS by ACP until an 18-carbon polyketide (dihydromonacolin L acid) is formed, being released by a thioesterase (TE) domain. The steps from the varying carbon processing cycles carried out by a methyl transferase (ME) binding a S-adenosyl methionine cofactor (SAM), ketoreductase (KR), dehydration (DH) and/or enoyl reductase (ER) domains are shown. The intermediate of each cycle is represented in the bottom left box. Mechanisms based on canonical proposals (such as those by MAT and TE) do not include residue numbers

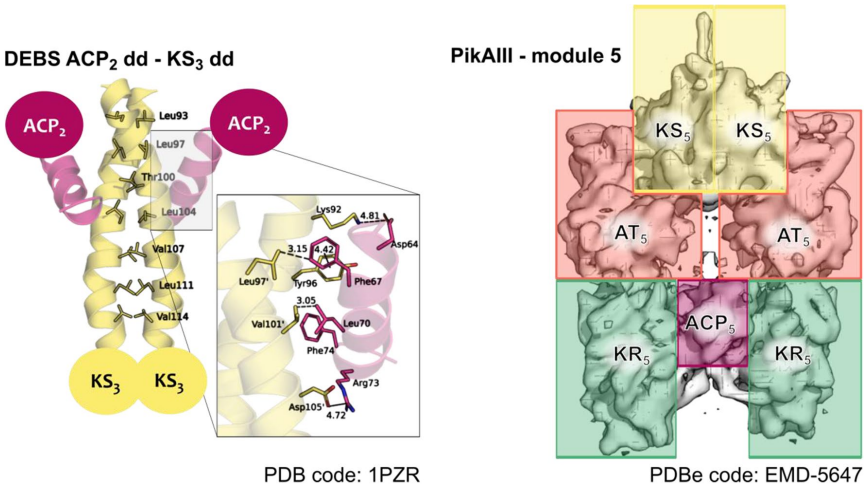


Fig. 3 On the left, a representation of the interacting *C*- and *N*-terminal docking domains (dd) of modules II and III of the DEBS PKS that promote intermodular non-covalent contacts, with the most relevant interactions highlighted [38]. On the right, a representation of the module 5 dimer of the pikromycin PKS PikAIII, from the pikromycin biosynthetic pathway [14]

building blocks without resorting to ACP to carry the growing polyketide [41].

2.1 Type-I PKS

Type-I PKS systems, schematically depicted in Figs. 4, 7, constitute the most complex among all PKS system types [1]. They can function as large multi-modular PKS assemblies or iteratively, with multiple elongation steps within a single cycle. Each module of the multi-modular PKS assemblies specifically processes a single elongation step of the polyketide synthesis, and therefore the necessary domains for the whole biosynthesis are simply replicated the number of times they are needed for the overall cycle, as opposed to the iterative architectures, in which the same module must act multiple times. As a result, and depending on domain assembling, type-I PKS systems can synthesize saturated, unsaturated and aromatic, linear and cyclic polyketides [2]. Moreover, they can form also "dynamic hybrid complexes" either with individual enzymes of the type-II PKS system or with nonribossomal peptide megasynthases (NRPSs), as exemplified for the equisetin biosynthesis in Fig. 4.

Type-I iterative PKSs, mostly found in fungi, synthesize almost every type of polyketides [1, 8, 42], whereas bacteria exhibit mostly a modular organization with few iterative modules found to date [1, 43]. In particular for type-I iterative PKSs, a phylogenetic analysis of the class in fungi showed that groups of genes encoding iterative PKS systems fall into eight groups: four encompassing reducing PKSs and four others encompassing non-reducing PKSs [44]. Some representative examples for non-reducing, partially-reducing and highly-reducing iterative PKSs are depicted in Fig. 4. The polyketide synthesis by type-I iterative PKSs resembles that of the fatty acid synthesis by animal FAS (Figs. 4, 5). The ACP domain binds a starting unit from the AT domain and loads it into the KS domain, and then repeats the process with extender units (generally malonyl-CoA) to carry out polyketide elongation. At each elongation cycle, the extender units are selectively carried by the ACP domain to carbon processing domains, such as the KR, DH or ER domains found in FAS, depending on the domain composition of each PKS. The resulting polyketides can either be released to the bulk by a TE domain (refer to the biosynthesis of dihydromonacolin L acid by the lovastatin PKS in Fig. 4), or transferred to other megasynthases by other protein carriers, where they will undergo different elongation cycles by specific domains (refer to equisetin biosynthesis in Fig. 5 for an example where the PKS counterpart of equisetin synthase synthesizes an octaketide precursor, which is then transferred to a peptidyl-carrier protein from the NRPS module and released by its Dieckmann cyclase).

As can be seen in Fig. 6, the highly-reducing iterative PKSs are very close to animal FAS in terms of domain organization, as they contain ACP, AT, KS, DH and ER domains identical to animal FAS, as well as the ME domain, similar to the non-catalytic ψME domain of animal FAS

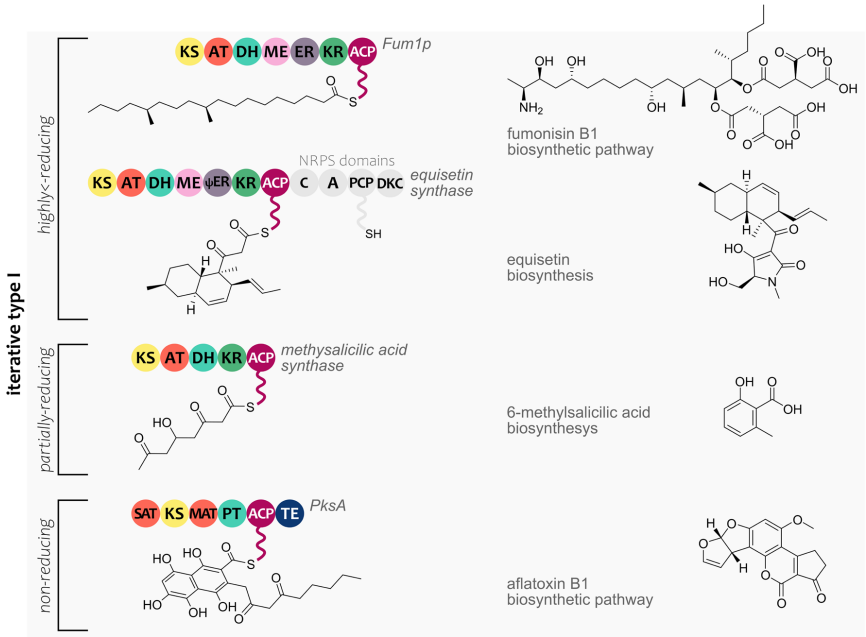


Fig. 4 Schematic representation of iterative type-I PKSs, highlighting representative biosynthetic pathways from highly-reducing to non-reducing iterative PKS. Domains are colored according to their similarity in function and structure: ketosynthases (KS) in yellow; acyl-CoA transferases (MAT, AT and SAT) in red; ketoreductases (KR) in forest green; dehydratases (DH), and product transferases (PT) in turquoise green; enoyl reductases (ER, ψER) in dark violet; methyltransferases (ME) in pink; acyl-carrier proteins (ACP) in dark magenta; and thioesterases (TE) in navy blue. Domains represented in light gray represent condensation (C), adenylation (A), peptidyl-carrier proteins (PCP) and Dieckmann cyclase (DKC) domains present in non-ribosomal polypeptide synthases that can be found in so-called hybrid-PKSs (e.g., equisetin synthase)

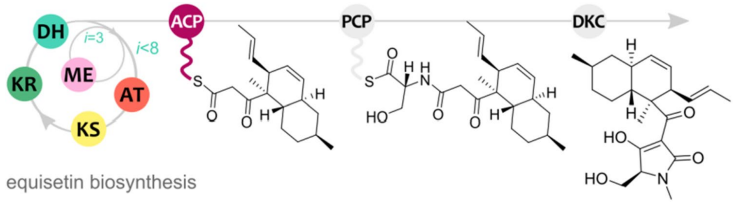


Fig. 5 Schematic representation of the equisetin biosynthesis by equisetin synthase, comprising a highly-reducing type-I iterative PKS and a non-ribossomal polypeptide synthase module. Domains of the PKS counterpart are colored according to their function: ketosynthases (KS) in yellow; acyl-CoA transferases (AT) in red; ketoreductases (KR) in forest green; dehydratases (DH); methyltransferases (ME) in pink; and acyl-carrier proteins (ACP) in dark magenta. Domains of the non-ribosomal polypeptide synthase module, peptidyl-carrier proteins (PCP) and Dieckmann cyclase (DKC), are represented in light gray

[45–47]. Consequently, polyketide products can potentially adopt all possible carbon oxidation states (between the initial acetyl and the final C-C saturated bond) under proper programming of the abovementioned catalytic domains, namely by site-directed mutagenesis to alter substrate specificity at a given iteration. Experiments seem to pinpoint the carbon processing domains as the most likely source of variability for the mechanism of highly-reducing iterative PKSs,

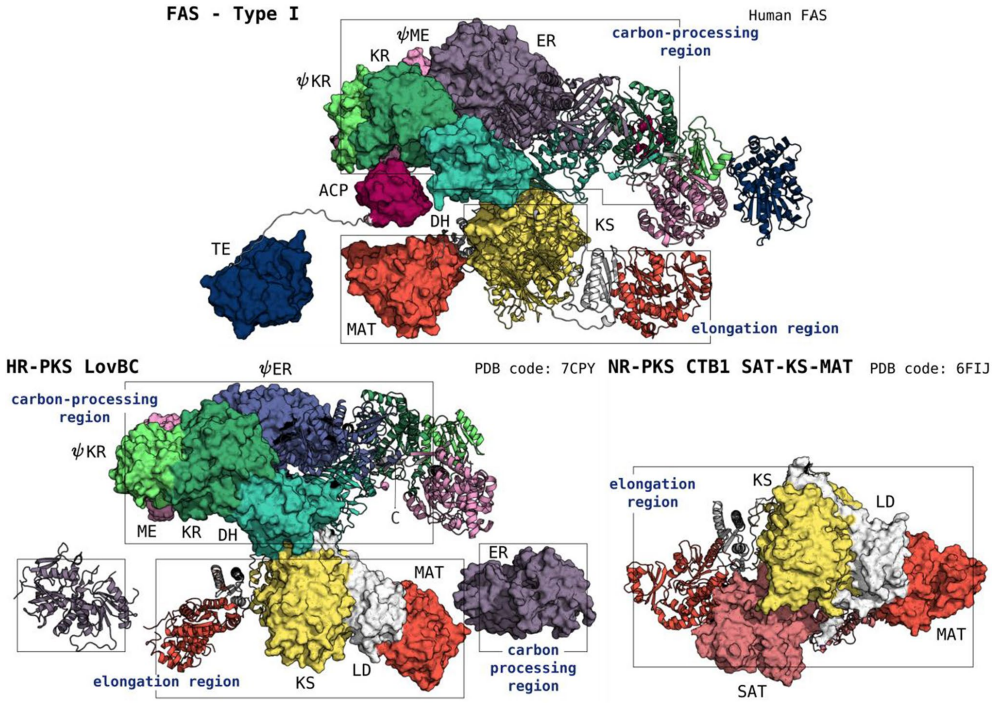


Fig. 6 Three dimensional representation of the human FAS (top panel) [45, 50], the highly-reducing PKS (HR-PKS) lovastatin nonaketide synthase LovBC complex (bottom left panel)[17] and the loading/elongation region of the non-reducing PKS (NR-PKS) CTB1 (bottom right panel) [16]. Within each panel, the monomer pairs are represented in cartoon (chain A) and surface (chain B). The overall architectures of the human FAS and the LovBC PKS exhibit an identical X-shaped structure, in which the domains that form the carbon-processing region (except the ER of the HR-PKS LovBC complex) occupy the top part of the "X", while those that comprise the elongation region form the bottom portion. The depicted CTB1 comprises only the domains that form the elongation region of this PKS, which also shares structural similarities with the elongation regions of the remaining two represented systems. Domains are colored according to their similarity in function and structure: ketosynthases (KS) in yellow; acyl-CoA transferases (MAT and SAT) in red; ketoreductases (KR, ψKR) in forest green; dehydratases (DH) in turquoise green; enoyl reductases (ER, ψER) in dark violet; methyltransferases (ME, ψME) in pink; acyl-carrier proteins (ACP) in dark magenta; thioesterases (TE) in navy blue; and the linker domain (LD) is colored in light gray. The condensation domain (C) present in LovBC is colored in dark gray

unlike type-II or type-III PKSs, in which the polyketide growth relies mostly on substrate specificity of the KS or AT domains. In particular, the ME, KR or ER domains were identified as relevant for polyketide variability in different pathways [48, 49]. Hence, the final PKS efficiency should be probably driven by kinetic competition of most iterative PKS domains with ACP-bound intermediates.

On the other hand, non-reducing iterative type-I PKSs, such as the CTB1 in Fig. 6, lack the KR, DH and ER domains. Hence, similarly to type-II and type-III PKSs, highly reactive polyketides are formed at each condensation cycle, by altering the substrate specificity of the KS domain towards longer polyketides [51, 52], and aromatic polyketides are favored as end products. More interestingly, while KSs are the starting domains of PKSs, the *N*-terminus of non-reducing iterative PKSs was predicted by the 3D-PSSM Protein Fold Recognition (Threading) Server to exhibit unexpected structural similarity to MAT domains and further sequence analysis revealed a GX**C**XG motif matching the GH**S**XG motif binding covalently to the malonyl extenders in MAT [53]. Since the TE domain is often absent in highly-reducing iterative PKSs (such as in Fum1p and equisetin synthase illustrated in Fig. 4) and given that, opposite to highly-reducing iterative PKSs, the AT domain in non-reducing iterative PKSs is specific towards malonyl-CoA, the identified starter ACP transacylase unit (SAT) of

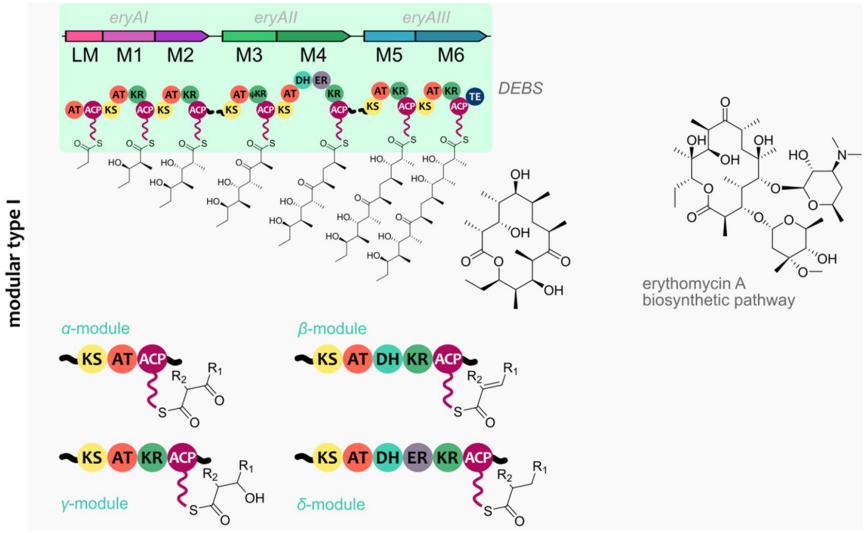


Fig. 7 Schematic representation of the prototypical DEBS modular type-I PKS. Domains are colored according to their function: ketosynthases (KS) in yellow; acyl-CoA transferases (AT) in red; ketoreductases (KR) in forest green; dehydratases (DH), enoyl reductases (ER) in dark violet; acyl-carrier proteins (ACP) in dark magenta; and thioesterases (TE) in navy blue. The gene organization of the modular PKS DEBS is shown (*eryAI*, *aryAII*, *eryAIII*) alongside the encoded modules and their domain organization

non-reducing iterative PKSs is thought to catalyze the transfer of reduced starters from the ACP of highly-reducing iterative PKSs (and even some animal FASs) to the non-reducing iterative PKSs [2]. Then, to ensure controlled product aldol cyclization and aromatization by non-reducing iterative PKSs, a transfer domain (PT) sharing structural similarities with the DH domain [54] has been always observed in this particular class. The PT domain is also responsible for assisting the product transfer from the KS domain to the TE domain (refer to PksA in Fig. 4) [51]. Finally, aside from the Claisen cyclization and hydrolysis reactions that are common in type-II PKSs, several TE domains of fungi iterative PKSs can promote product release by intracyclization via Dieckmann condensation, instead of the common hydrolysis reaction [55].

Understanding evolution, function and structure of type-I PKSs becomes increasingly more complex as modular PKSs are concerned. Instead of working iteratively, modular PKSs resemble an assembly of iterative PKSs that work non-iteratively to produce various polyketides. The assembling of PKS modules can occur either covalently, when the modules are encoded in a single gene, or through protein:protein interactions by terminal docking domains, where a proper orientation between modules is promoted [38]. In Fig. 7 the modular organization of the prototypical DEBS modular PKS is illustrated.

Each module initiates with the KS and AT domains required for polyketide elongation, followed by the required carbon processing domains (α-, β-, γ- and δ-modules in Fig. 7), and terminates with the ACP domain required to shuttle the final polyketide from one module to another. Short molecular dynamics (MD) simulations of a KS-AT didomain from a minimal PKS module (α-module in Fig. 7) further emphasized the role of ACP in substrate shuttling, since small inter- and intradomain rearrangements were observed for the KS-AT didomain regardless of a 100-long amino acid linker region [56]. A few modular PKSs that do not include the AT domain in their modules, relying instead in a single AT domain to feed the extender ketide unit, were also characterized. As a result, the question of whether modular PKSs have evolved from gene duplication and recombination events of iterative PKSs or are simply the result of more complex recombination events remains open [57].

The interaction between some catalytic domains, namely KR or DH, and the ACP domain carrying the polyketide substrate were also explored with docking calculations and seminal MD simulations in a PKS module of the erythromycin synthase [58]. The simulations resulted in polyketide binding modes that were suggested to resemble those of other reductive PKSs: while polyketide binding to the KR domain was observed to occur in a shallow exposed cavity close to residues known to be involved

in stereospecificity [59], the polyketide was observed to occupy a buried tunnel in DH, resembling the binding of a type-II FAS inhibitor of *E. coli* [60]. Seminal calculations combining docking and MD simulations on a KR domain of the amphotericin PKS in complex with two stereoisomers, were also able to further identify residues at the active site that could help rationalize its stereospecificity [61]. In addition, the mechanism of dehydration and isomerization of a DH domain from gephyronic acid PKS was determined alongside important residues for substrate recognition, using a combination of QM/MM calculations and molecular mechanics/generalized Born surface area (MM/GBSA) calculations [62]. Even though these results provided a glimpse at the complexity of the catalytic machinery of PKSs and their products, the availability of atomistic depictions of their function remains scarce.

Analogously to type-II PKSs and type-I non reducing iterative PKSs, the final module also includes a TE domain for product release by either intracyclization or hydrolysis. The reaction mechanism of the TE domain of modular PKSs has been studied the most with hybrid QM/MM simulations, in an attempt to rationalize the occurrence of either the hydrolysis or macrocyclization reactions [63–65]. Rather than the common hydrolysis step observed in the TE domain of FASs, MD simulations of the TE domain located in the last module of the DEBS prototypical type-I PKS suggested that a hand-shaped pocket of hydrophobic residues, comprised by Ala77, Leu183, Leu186, Leu190 and Phe260, interacted strongly with the polyketide substrate, thus preventing access of water to the active site of TE and favoring a pre-reactant state by orienting the polyketide-hydroxyl towards the His-Asp composing the conserved Ser-His-Asp triad [64]. A similar binding was also observed for the TE domain of pikromycin [63], where it was proposed to favor the orientation of the farthest hydroxyl of the substrate towards the catalytic His, consequently placing it nearby the catalytic Ser binding covalently to the polyketide substrate, which, as confirmed by QM/MM calculations, made the cyclization reaction feasible upon overcoming energy barriers around 20 kcal·mol^{-1} [64, 65].

2.2 Type-II PKS

To date, type-II PKSs were only found in bacteria [1]. In the starting and condensations steps, the ACP domain carries the starter and the extender units from the MAT enzyme to KS, where polyketide elongation occurs. Figure 8 shows a schematic representation of the organization and mode of action of type-II PKS enzymes. Apart from the promiscuity of the KS enzymes of this class towards starter units for polyketide synthesis, the three main differences between type-I and type-II PKSs are presented in more detail.

The first difference concerns the assembly of type-II PKSs. Type-II PKS KS assemble in heterodimeric units, where one unit (KS_α) conserves the Cys-His-His catalytic motif (Fig. 9, top left) and performs the Claisen-condensation, while the second unit (KS_β) is actively involved in the control of polyketide length (Fig. 9, bottom), as suggested by quantification and characterization of polyketide products from mutated and chimeric PKS-KS heterodimers [66–68]. Based on amphiphilic nature of the binding pocket in KS_β, as revealed in the sole X-ray structure of the KS heterodimer [19], it has been suggested that several rounds of Claisen-condensation occur in KS before an aldol condensation releases the polyketide from KS. However, modeling of linear and cyclic polyketides through atom replacement suggested that ACP interacts strongly with linear polyketides [69], which poses contradictory evidence as to the occurrence of the aldol condensation in KS. A study combining the X-ray characterization and QM/MM simulations of the minimal type-II PKS system responsible for the anthroquinone biosynthesis also showed that there is a narrow cleft in KS_α:KS_β capable of allocating the growing polyketide and prevent spontaneous cyclization [70]. Despite that the type-II KS is a heterodimer, this study emphasized that several active site residues, also conserved in homodimeric KS from both FAS and PKS, were clearly involved in polyketide binding, namely the Cys-His-His triad and a Phe (Fig. 9, top) proposed to stabilize an oxyanion formed upon the Claisen condensation step [70, 71].

The second difference has to do with the limited set of reactions catalyzed by type-II PKSs as opposed to type-I

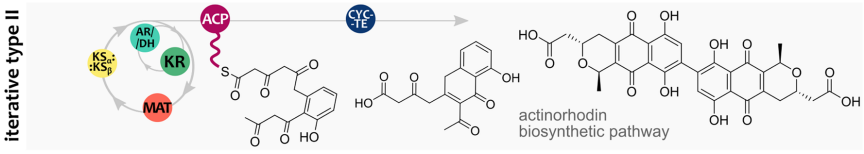


Fig. 8 Schematic representation of the type-II PKSs through the most studied actinorhodin biosynthetic pathway. Domains are colored according to their function: ketosynthase (KS_α:KS_β) in yellow; acyl-CoA transferase (MAT) in red; ketoreductase (KR) in forest green; dehydratase (DH) and aromatase (AR) in turquoise green; acyl-carrier proteins (ACP) in dark magenta; and cyclase-thioesterase (CIC-TE) in navy blue

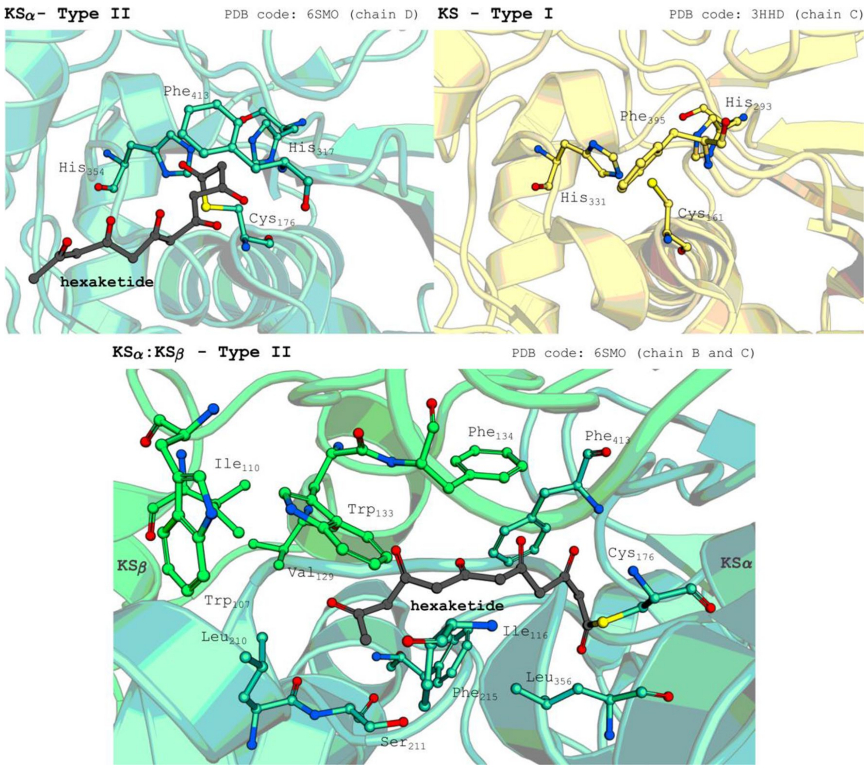


Fig. 9 Comparison of the active site of the ketosynthase (KS) of a type-II PKS involved in anthraquinone biosynthesis complexed with a hexaketide substrate (top left) and the KS domain of the animal FAS (top right). Conserved residues are represented on the upper panels in ball-and-stick representation, namely the catalytic triad Cys-His-His and the Phe switch contributing for an oxyanion hole [70, 71]. The binding of the hexaketide at the interface of the heterodimeric type-II KS_{α}:KS_{β} is depicted in the bottom panel (residues from KS_{α} are represented in pale green ball-and-sticks and those of KS_{β} are represented in bright green) [70]

PKSs. Only KR and, possibly, DH enzymes were identified in the type-II PKS systems, which means that, in light of current knowledge, type-II PKSs do not catalyze saturated C–C bond formation and, consequently, their main products are unsaturated and aromatic polyketides. On a side note, the detection of type-II PKS-DHs has only recently been carried out in three pathways synthesizing polyene precursors [72–74]. Characterization of polyketide products from recombinant type-II PKSs showed that most type-II PKS KR (Fig. 10) catalyze polyketide cyclization with a high regio- and stereoselectivity [75, 76]. As a result, type-II PKS KR are particularly important enzymes in the differentiation of the polyketide products of type-II PKSs [77, 78]. Further structural characterization of type-II PKS KR enzymes, by combined x-ray crystallography and molecular docking, suggested that the specific cyclization by KR is a result of both polyketide binding and the quaternary structure of the KR tetramer [18], and it has identified key residues for the tuning of such specificity [77].

More recently, MD and free-energy simulations of the type-II PKS KR involved in the actinorhodin biosynthesis, with three different substrates, revealed that the stereoselectivity of the enzyme depends both on the substrate binding pose and the relative kinetic rate of the *R*/*S* -enantiomers. With the study of the wild-type and mutant forms of enzyme, the authors concluded that, while *S*-enantiomers were kinetically favored, the preference for *R*-enantiomers depended on the binding energy of the substrate to the active site of the enzyme [79]. Furthermore, it was also concluded that one can intentionally control the enzyme's stereoselectivity by modulating its catalytic efficiency (increasing/decreasing

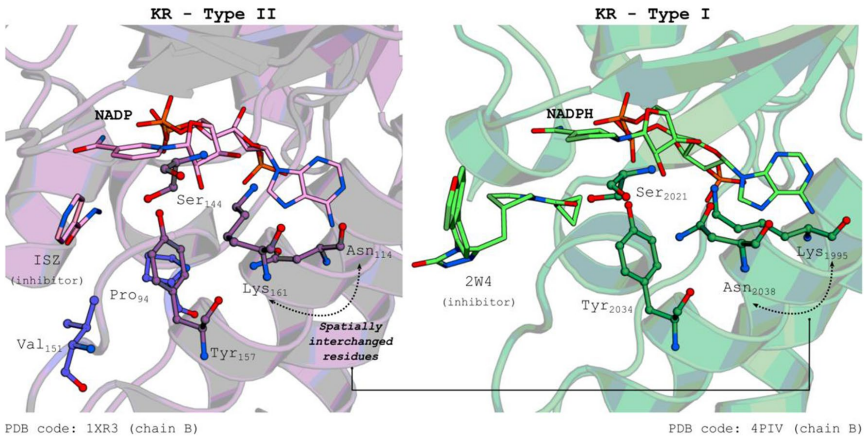


Fig. 10 Comparison of the active site of the ketoreductase (KR) of a type-II PKS involved in actinorhodin biosynthesis complexed with the 4-(diazenylcarbonyl)pyridine inhibitor (ISZ) and the $NADP^+$ cofactor (left) and the KR domain of the animal FAS complexed with the GSK2194069 inhibitor (2W4) and the NADPH cofactor (right). Conserved residues are represented in ball-and-stick representation, namely the catalytic Ser and Tyr [18, 46]. The Pro94 and Val151 proposed by Serapian and coworkers to control the stereospecificity of KR in actinorhodin biosynthesis are also highlighted in dark blue [79]

the stabilization of the transition-state structure) and/or by redesigning the substrate's orientation in the active site.

The third difference comes down to the process behind the release of the final polyketide product. Contrary to type-I PKSs, the synthesis of the final aromatic polyketide products by type-II PKSs depends on the activity of a type-II PKS AR enzyme and of a final TE enzyme that functions both as a TE and a cyclase (CYC-TE). While the first catalyzes the dehydration of the cyclic type-II PKS KR product [80], the latter is responsible for the final product cyclization and release. [20, 80]

2.3 Type-III PKS

Most recent phylogenetic studies indicate that the distinct type-III PKSs are found in all land plants and a few lineages of bacteria and fungi [81]. Unlike other PKSs or FASs, type-III PKS systems do not rely on ACP to perform acyl-CoA supply to KS enzymes. Instead, they are KS homodimers that bind different CoA starters, ranging short acyl-CoA (up to 6-carbon long), large acyl-CoA and aromatic acyl-CoA (e.g., benzoyl-CoA [82] or *p*-coumaroyl-CoA [83, 84]), and catalyze a number of successive condensation reactions of malonyl-based or acetyl blocks [85, 86], based on the size of their active site cavity. The resulting products are highly oxidant compounds that are prone to cyclization by either Claisen or aldol condensation and lactonization, after tetraketide formation [87], and often constitute important precursors in the synthesis of functional compounds involved in protection, resistance or differentiation mechanisms [81].

The general reaction types of type-III PKSs can be described as a combination of three reaction steps: (i) loading of a starter acyl-CoA substrate, (ii) elongation of the starter acyl-CoA with the extender acyl-CoA through a Claisen condensation, (iii) cyclization of the polyketide (refer to Fig. 11 as an example). Through specific combinations of these reactions, a total of about 53 different reaction types were predicted for type-III PKSs; however, to date, only 15 were observed [81]. Structure analysis and mutagenesis studies of type-III PKSs have identified residues contributing for substrate specificity, the condensation reactions or the types of cyclization undergoing in the polyketide product [81, 87]. In a recent computational study, QM/MM calculations suggested that the relative positioning of a Phe contributed for the shape of the binding pocket of type-III PKS chalcone synthases (CHS) and stilbene synthases (STS), and the subsequent aldol or Claisen cyclizations taking place. Similarly to the role of Phe395 in animal FAS, which regulates the access of the substrate to the active site of KS [88], the Phe in STS favored the interaction of the tetraketide substrate with a larger cyclization site, where an aldol condensation could occur, while simultaneously decreased the binding pocket of the enzyme relatively to that of CHS [89]. Taking into account water-mediated mechanisms including a Thr-Arg dyad in CHS and a Thr-Glu dyad in STS, the

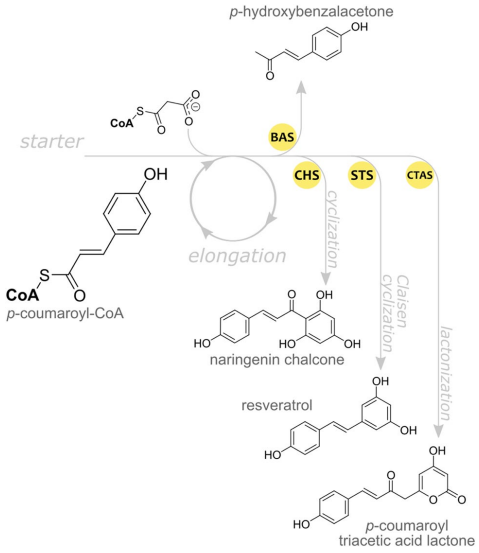


Fig. 11 Schematic representation of the possible transformations performed by different type-III PKSs, using *p*-coumaroyl-CoA as a starter unit. Upon elongation cycles by type-III PKSs with malonyl-CoA, different products are obtained: *p*-hydroxybenzalacetone is obtained after Claisen condensation by benzalacetone synthases (BAS), resveratrol is obtained after Claisen condensation by stilbene synthases (STS), naringenin chalcone is obtained after a cyclization by chalcone synthases (CHS), and *p*-coumaroyl triacetic acid lactone is obtained after lactonization by *p*-coumaroyl triacetic acid synthase (CTAS). The corresponding type-III PKS (BAS, CHS, STS, CTAS) are colored in yellow

authors proposed an unified mechanism for the enol-keto tautomerization taking place during cyclization by these enzymes [89]. Mutagenesis studies on type-III PKS CHSs and STSs have also shown that site-directed mutagenesis could be employed to favor either Claisen or aldol cyclization, and consequently interconvert the products of type-III PKSs, even though with reduced final yields [87, 90, 91]. As a result, bioinformatic tools combining phylogenetic analysis, sequence alignments and structural data were made available to predict the reaction type of type-III PKSs based on novel sequence alignments [92–94]. Since few structural data on type-III PKSs is available to date, new promising developments in this field should be expected as more type-III PKSs will be characterized.

3 Engineering of PKSs

The PKS megaenzymes are examples of complex nanotechnological machines, with enormous potential to biosynthesize high-value active pharmaceuticals, agrochemicals, biofuels and industrial chemicals, among others. Recent years have witnessed the engineering of PKSs through different techniques, the main ones briefly described subsequently.

3.1 Module Swapping

Some of the more common strategies for PKS engineering focus at engineering synthetic PKS genes through module swapping, in order to produce varying chimeric PKSs and seek characterization of resulting products [95, 96].

Several engineered modular PKSs and hybrid modular PKSs (obtained through module swapping with NRPS megasynthases) have been shown to produce new polyketides following the co-linearity principle, where molecular synthesis is driven by the vectorial nature of the modular organization [6, 97, 98]. However, the yield of the engineered module assemblies was often lower than that of native modular PKSs [99–101]. Early kinetic studies pointed out that linker regions in between modules were required for modular PKS activity [102–104], which is in line with recent structural insight on the role of covalent linkers and *C*-/*N*-termini module docking domains in module communication [14, 15, 38]. These so-called docking domains, found at the start of the downstream module and at the end of the upstream module, were shown to possess a high helical content that contributed for intermodule docking and dimerization, and consequently to PKS efficiency. However, in vitro kinetic studies on the prototypical DEBS modular PKSs have also indicated that communication alone does not explain the overall efficiency of PKSs, as a bimodular derivative was observed to exhibit a slower maximum turnover than a trimodular one or the complete hexamodular DEBS [105].

More recently, engineering of modular PKSs from repetitive gene splicing junctions in rapamycin and pikromycin biosynthetic gene clusters delivered highly productive assembly lines [96, 98]. These studies identified splicing junctions often located within AT and KS sequences or linker regions upstream of ACP. Together with phylogenetic studies, further heterologous expression of several polyketide gene clusters has put out a hypothesis that module boundaries in modular PKSs may be established from the upstream AT to the downstream KS domain, instead of the conventional KS to ACP order observed in iterative PKSs and in animal FAS [106].

As examples of successful PKS engineering[3, 96] and structural characterization of PKS module assembling[14, 15] become more frequent in literature, the innovations in bioinformatic tools and phylogenetic analysis algorithms, as well as further in vitro characterization of novel modular PKSs and their reaction products, will surely provide great advancements in the field.

3.2 Domain Swapping

Attempts to explore bacterial modular PKSs through domain swapping are less frequent in literature [3, 9]. In particular, the AT domains are preferred targets for domain swapping, as they are primary responsible for polyketide diversification by changing the nature of the acyl extender, and they are the more promiscuous out of all PKS domains [107]. However, AT domain swaps in modular PKSs often led to loss of enzyme efficiency or even unproductive modular PKSs, due either to improper gene splicing or inefficient domain communication [108, 109]. KR swaps made to sculpt polyketide stereochemistry have been also carried out with some success [110], although there were not many attempts carried in highly reductive modules, nor clear trends between KR domain swapping and module functionality could be observed [3, 107].

Instead of the more cumbersome domain deletion, inactivation of specific domains by mutagenesis has also been suggested for control of the polyketide reductive cycle [107].

3.3 Directed Engineering at Specific Amino Acid Positions

A very common strategy for enzyme engineering is the use of directed engineering at specific amino acid positions. The main advantage of specific amino acid mutations is the possibility to alter substrate specificity and the catalytic efficiency of particular domains, while keeping most protein:protein interactions required to produce a competent modular PKSs [111]. This approach has been employed in the engineering of type-III PKSs for the production of aromatic polyketide products, because type-III PKS function as homodimers, not requiring additional enzymes for polyketide synthesis.[112]

A particularly relevant example is the case of KS domains, which were observed to exhibit poor kinetics towards unnatural substrates [113]. By inspecting the sequence of a KS domain from erythromycin polyketide synthase and that of mycolactone polyketide synthase, site-directed mutagenesis carried out in the former in vitro, improved its rate of condensation against several substrates [114]. AT domains, which exhibit higher promiscuity among other PKS domains [115, 116], have also been often subject to site-directed mutagenesis studies focusing on the metabolic engineering of polyketide pathways [107, 111, 117, 118].

3.4 First-Line Antibiotics, Anticancer Drugs, Veterinary Medicines and Agrochemicals

Since PKSs can synthesize a wide range of complex bioactive polyketide natural products, namely environmental toxins, antibiotics, antineoplastics or immunosuppressants, from simple acyl-CoA building units, the last decades witnessed an increasing attention for the exploration of the potential of the PKSs for in the de novo synthesis of novel compounds with applicability in health and technology.

In particular, type-II PKSs are highly attractive enzymatic systems and the characterization of their function and specificity (in particular for the actinorhodin biosynthesis pathway of *Streptomyces coelicolor*) [4] was crucial to outline a set of strategies for the rational design of polyketides, as type-II PKSs synthesize a large family of highly functionalized polycyclic aromatic compounds, including some first-line antibiotics, anticancer drugs, veterinary medicines and agrochemicals [7, 119, 120]. Focusing on the possibility, despite uncommon, to accept different starter units and on the unique properties of the heterodimeric KS for chain elongation of polyketides, different aromatic polyketides were synthesized [121]. As an example, propionyl-CoA was observed to be the starter unit of lomaiviticins [122], a class of compounds with antibiotic and antitumor activities, and daunorubicin and doxorubicin [123, 124], both commercial chemotherapeutical drugs. The diversity of produced polyketides can be further expanded upon inclusion of modified amino acid starter units, which are particularly profitable for the synthesis of natural antibiotics [125]. The unique ability of the heterodimeric KS_{α}:KS_{β} complexes to control polyketide chain elongation has also been exploited to synthesize novel antibacterial or antitumor compounds, namely huanglongmycins and saccharothrixones [126, 127]. Finally, the high degree of functionalization and aromaticity renders the resulting polyketides as good reagents for tailoring reaction by several oxygenases, broadening the spectrum of products and their applicability [121].

3.5 New Polyketide Pathways and the Synthesis of New Products

Some recent reports of type-I PKS engineering also include the synthesis of biofuels and industrial chemicals from PKS products [128–130], even though current efforts consistently report low product yields (below 100 $mg \cdot L^{-1}$) due to their frequent slow turnover rate, metabolic burden or product toxicity [131]. Nevertheless, the development of novel chemical biosynthetic routes by engineering modular type-I PKSs is quite an enticing topic of research. The fact

that each module of the PKS acts as in a straightforward assembly line (the so-called vectorial feature), has raised attention to this class as a means to identify and analyze new polyketide pathways, and as a rationale for PKS engineering to synthesize new products [3, 131]. In this regard, bioinformatics and chemoinformatics can constitute promising studies toward the rational engineering of these systems. The increasing availability of genetic and structural characterization of PKSs has propelled the bioinformatic tool ClusterCAD, which makes use of curated data on modular PKS gene clusters, protein sequence and resulting product, to predict chimeric PKS assemblies that may synthesize a polyketide of interest [132]. More recently, a chemoinformatic approach was devised in which module reductive loops were identified for PKS engineering by module swapping, based on the structural similarity between their resulting polyketide intermediates and the polyketide of interest. Upon the expression of designed chimeric lipomycin PKS modules in *Streptomyces albus* J1074, promising titers above 100 $mg{\cdot}L^{-1}$ for the synthesis of short-chain fatty acids and triketide lactones were observed [133].

Despite being in its infancy, rational engineering strategies exploiting the biosynthetic potential of the PKS family of enzymes to produce polyketides with wide use in the pharma and chemical industries has held promising results. Hence, we can expect that the increasing identification of PKS-encoding genes and the structural characterization of PKS enzymes, as well as their products, will considerably increase the number of successful attempts at synthesizing customized polyketides.

3.6 Future Directions for Computational Chemists

All type-I PKS, both modular and iterative ones, are homodimeric megaenzyme assemblies with a molecular weight of several MDa per monomer (often far larger than the largest FAS megaenzymes) [2, 3]. Hence, it is very challenging to perform simulations beyond the microsecond timescale that would be required to observe domain-domain motions contributing to polyketide synthesis. Despite that, a few computational studies have attempted to underly the rationale of intra- and interdomain motions in selected PKS modules; these comprised modest nanosecond MD simulations, and thus cannot infer such motions on a significant timescale [56, 58]. In particular for modular PKSs, a dynamic characterization of intermodular binding, namely by the *N*- and *C*-docking domains, could substantially contribute for improved engineering strategies, as recent studies suggest that they may be essential for the design of functional modular PKSs [96, 98]. In addition, the study of polyketide transport in both iterative and modular type-I PKSs may also contribute to unveil domain:domain interactions contributing to substrate binding, namely in what concerns the interaction of the ACP domain and the catalytic domains of PKSs. As for modular PKSs, current computational studies employed MD simulations of a few nanosecond to explore such interactions [50, 58, 61, 134], but only extensive MD simulations can provide more detailed information for further PKS engineering [135, 136]. In fact, recent computational mutagenesis validated by MD simulations and in vitro assays could identify around 10 residues within the active site of an AT domain of DEBS (M6 in Fig. 7) that improved its affinity towards provided extender units [118]. Combined docking and MD simulations of AT:substrate complexes of the modular salinomycin synthase also revealed hydrophobic residues at the active site of each that, upon mutation, could alter the specificity of the AT domains [137]. We expect that with the increasing availability of modular PKS structures from X-ray and cryo-EM studies, as well as novel MD sampling techniques and the increasing availability of computing resources, it will soon be possible to simulate the vectorial dynamics of type I modular PKSs, namely the intermodular transfer of the polyketide substrate along the different modules and its dimerization.

Since site-directed mutagenesis is also a popular option for PKS engineering [10, 111], the study of enzyme reaction mechanisms of individual PKS domains can greatly contribute for rational enzyme engineering strategies [138, 139]. In that regard, in recent years, we have characterized the reaction mechanisms of the MAT, KR, DH and ER domains of animal FAS with hybrid QM/MM methods [24–28]. Despite that these domains differ from those of PKSs, their catalytic residues and core reaction mechanisms resemble each other. Hence, it might be interesting if future studies address the differences in the substrate specificity between the enzymes/domains of FASs and PKSs, namely within their active sites. The ability of TE domains in modular PKSs to conduct either polyketide macrocyclization or hydrolysis has also been explored with mechanistic detail using hybrid QM/MM simulations [63–65]. On the other hand, mechanistic studies of standalone PKS enzyme (type-II and type-III) are quite scarce [79, 89], and would be beneficial to explore the diversity of reactions that are possible by these systems, namely the Claisen condensation, aldolization or lactonization reactions performed by KSs of type-III PKSs [81, 112], or the stereoselective cyclization and ketone reduction, or the dehydration and aromatization reactions catalyzed by KRs and DHs of type-II PKSs [4, 121]. In addition, there is a large possibility of substrates or even extender units for standalone PKSs, which render the modelling of competent enzyme:substrate complexes even more difficult, and an accurate mechanistic study rather demanding.

In light of the recent advances and the wide potential of PKS bioengineering, the years to come will surely offer insightful evidence from computational chemists, which will propel rational engineering strategies to assemble efficient

PKSs capable of synthesizing products with broad societal impact.

4 Conclusions

The PKS megaenzymes are examples of extraordinary nanotechnological machines, with a huge potential to biosynthesize high-value active pharmaceuticals, agrochemicals, biofuels and industrial chemicals, among others. Indeed, PKSs can perform sophisticated chemistry, accommodating different substrates yet synthesizing the desired products with accuracy. We pointed out the current structural, biochemical and catalytical knowledge on these megaenzymes, even though we were not able to fulfill the existing void on their complex catalytic pathways, which still present a challenge to the scientific community. Additionally, the significant resemblance between the PKS and the FAS megaenzymes seems to suggest that there is some sort of parallel that we can set between both systems, even though the first seem to be more complex and more promising than the latter, as far as the engineering of new products is concerned. However, all these amazing megasystems still require extensive research before they can efficiently achieve their synthetic potential. In that regard, even though a great deal of computational work has been carried out on the FAS megaenzymes, this is not the case for the PKS megaenzymes, which remain an open field for the computational chemist to explore.

Acknowledgements This work was supported by UIBD/50006/2020, PTDC/QUI-QFI/28714/2017 and PTDC/QUI-OUT/1401/2020 with funding from FCT/MCTES through national funds. PP, MV, PF and JPMS thank FCT/MCTES (Fundação para a Ciência e Tecnologia/Ministério da Ciência, Tecnologia e Ensino Superior) and ESF (European Social Fund) through POCH (Programa Operacional Capital Humano) and PORN (Programa Operacional Regional do Norte) for their PhD Grants ref. PD/BD/135268/2017, SFRH/BD/119206/2016, SFRH/BD/144205/2019 and SFRH/BD/131968/2017, respectively. FEM thanks ANID FONDECYT 3200304. RPPN acknowledges funding from project PTDC/QUI-QFI/28714/2017 through FCT/MCTES.

Author Contributions RPPN, PF, FEM, PP, JPMS and MFV contributed equally to the work.

References

1. Jenke-Kodama H, Sandmann A, Muller R, Dittmann E (2005) Evolutionary implications of bacterial polyketide synthases. Mol Biol Evol 22(10):2027–2039. https://doi.org/10.1093/molbev/msi193
2. Herbst DA, Townsend CA, Maier T (2018) The architectures of iterative type I PKS and FAS. Nat Prod Rep 35(10):1046–1069. https://doi.org/10.1039/c8np00039e
3. Klaus M, Grininger M (2018) Engineering strategies for rational polyketide synthase design. Nat Prod Rep 35(10):1070–1081. https://doi.org/10.1039/c8np00030a
4. Chen AC, Re RN, Burkart MD (2018) Type II fatty acid and polyketide synthases: deciphering protein-protein and protein-substrate Interactions. Nat Prod Rep 35(10):1029–1045. https://doi.org/10.1039/c8np00040a
5. Staunton J, Weissman KJ (2001) Polyketide biosynthesis: a millennium review. Nat Prod Rep 18(4):380–416. https://doi.org/10.1039/a909079g
6. Cropp TA, Kim BS, Beck BJ, Yoon YJ, Sherman DH, Reynolds KA (2002) Recent developments in the production of novel polyketides by combinatorial biosynthesis. Biotechnol Genet Eng 19:159–172. https://doi.org/10.1080/02648725.2002.10648028
7. Das A, Khosla C (2009) Biosynthesis of aromatic polyketides in bacteria. Acc Chem Res 42(5):631–639. https://doi.org/10.1021/ar8002249
8. Weissman KJ (2010) Peering into the black box of fungal polyketide Biosynthesis. ChemBioChem 11(4):485–488. https://doi.org/10.1002/cbic.201000023
9. Poust S, Hagen A, Katz L, Keasling JD (2014) Narrowing the gap between the promise and reality of polyketide synthases as a synthetic biology platform. Curr Opin Biotechnol 30:32–39. https://doi.org/10.1016/j.copbio.2014.04.011
10. Stegemann F, Grininger M (2021) Transacylation kinetics in fatty acid and polyketide synthases and its sensitivity to point mutations. ChemCatChem 13:1–13. https://doi.org/10.1002/cctc.202002077
11. Cane DE, Walsh CT, Khosla C (1998) Biochemistry—harnessing the biosynthetic code: combinations, permutations, and mutations. Science 282(5386):63–68. https://doi.org/10.1126/science.282.5386.63
12. Moore BS, Hertweck C (2002) Biosynthesis and attachment of novel bacterial polyketide synthase starter units. Nat Prod Rep 19(1):70–99. https://doi.org/10.1039/b003939j
13. Zhou H, Zhan JX, Watanabe K, Xie XK, Tang Y (2008) A polyketide macrolactone synthase from the filamentous fungus Gibberella zeae. Proc Natl Acad Sci U S A 105(17):6249–6254. https://doi.org/10.1073/pnas.0800657105
14. Dutta S, Whicher JR, Hansen DA, Hale WA, Chemler JA, Congdon GR, Narayan ARH, Hakansson K, Sherman DH, Smith JL, Skiniotis G (2014) Structure of a modular polyketide synthase. Nature 510(7506):512. https://doi.org/10.1038/nature13423
15. Dorival J, Annaval T, Risser F, Collin S, Roblin P, Jacob C, Gruez A, Chagot B, Weissman KJ (2016) Characterization of intersubunit communication in the virginiamycin trans-acyl transferase polyketide synthase. J Am Chem Soc 138(12):4155–4167. https://doi.org/10.1021/jacs.5b13372
16. Herbst DA, Huitt-Roehl CR, Jakob RP, Kravetz JM, Storm PA, Alley JR, Townsend CA, Maier T (2018) The structural organization of substrate loading in iterative polyketide synthases. Nat Chem Biol 14(5):474. https://doi.org/10.1038/s41589-018-0026-3
17. Wang JL, Liang JD, Chen L, Zhang W, Kong LL, Peng C, Su C, Tang Y, Deng ZX, Wang ZJ (2021) Structural basis for the biosynthesis of lovastatin. Nat Commun 12(1):21174. https://doi.org/10.1038/s41467-021-21174-8
18. Korman TP, Hill JA, Vu TN, Tsai SC (2004) Structural analysis of actinorhodin polyketide ketoreductase: cofactor binding and substrate specificity. Biochemistry 43(46):14529–14538. https://doi.org/10.1021/bi048133a
19. Keatinge-Clay A, Maltby DA, Medzihradszky KF, Khosla C, Stroud RM (2004) An antibiotic factory caught in action. Nat Struct Mol Biol 11(9):888–893. https://doi.org/10.1038/nsmb808

20. Taguchi T, Awakawa T, Nishihara Y, Kawamura M, Ohnishi Y, Ichinose K (2017) Bifunctionality of ActIV as a cyclase-thioesterase revealed by in vitro reconstitution of actinorhodin biosynthesis in streptomyces coelicolor A3(2). ChemBioChem 18(3):316–323. https://doi.org/10.1002/cbic.201600589
21. Tsai SC, Ames BD (2009) Chapter 2 structural enzymology of polyketide synthases. In: Methods in enzymology, vol 459. Academic Press, pp 17–47. doi:https://doi.org/10.1016/S0076-6879(09)04602-3
22. Du LC, Lou LL (2010) PKS and NRPS release mechanisms. Nat Prod Rep 27(2):255–278. https://doi.org/10.1039/b912037h
23. Ames BD, Lee M-Y, Moody C, Zhang W, Tang Y, Tsai S-C (2011) Structural and biochemical characterization of ZhuI aromatase/cyclase from the R1128 polyketide pathway. Biochemistry 50(39):8392–8406. https://doi.org/10.1021/bi200593m
24. Paiva P, Sousa SF, Ramos MJ, Fernandes PA (2018) Understanding the catalytic machinery and the reaction pathway of the malonyl-acetyl transferase domain of human fatty acid synthase. ACS Catal 8(6):4860–4872. https://doi.org/10.1021/acscatal.8b00577
25. Paiva P, Sousa SF, Fernandes PA, Ramos MJ (2019) Human fatty acid synthase: a computational study of the transfer of the acyl moieties from MAT to the ACP domain. ChemCatChem 11(16):3853–3864. https://doi.org/10.1002/cctc.201900548
26. Medina FE, Neves RPP, Ramos MJ, Fernandes PA (2017) A QM/MM study of the reaction mechanism of human beta-ketoacyl reductase. Phys Chem Chem Phys 19(1):347–355. https://doi.org/10.1039/c6cp07014k
27. Medina FE, Neves RPP, Ramos MJ, Fernandes PA (2018) QM/MM study of the reaction mechanism of the dehydratase domain from mammalian fatty acid synthase. ACS Catal 8(11):10267–10278. https://doi.org/10.1021/acscatal.8b02616
28. Medina FE, Ramos MJ, Fernandes PA (2019) Complexities of the reaction mechanisms of CC double bond reduction in mammalian fatty acid synthase studied with quantum mechanics/molecular mechanics calculations. ACS Catal 9(12):11404–11412. https://doi.org/10.1021/acscatal.9b03531
29. Paiva P, Medina FE, Viegas M, Ferreira P, Neves RPP, Sousa JPM, Ramos MJ, Fernandes PA (2021) Animal fatty acid synthase: a chemical nanofactory. Chem Rev. https://doi.org/10.1021/acs.chemrev.1c00147
30. Ames BD, Nguyen C, Bruegger J, Smith P, Xu W, Ma S, Wong E, Wong S, Xie X, Li JWH, Vederas JC, Tang Y, Tsai S-C (2012) Crystal structure and biochemical studies of the trans-acting polyketide enoyl reductase LovC from lovastatin biosynthesis. Proc Natl Acad Sci. https://doi.org/10.1073/pnas.1113029109
31. Storm PA, Herbst DA, Maier T, Townsend CA (2017) Functional and structural analysis of programmed C-methylation in the biosynthesis of the fungal polyketide citrinin. Cell Chem Biol 24(3):316–325. https://doi.org/10.1016/j.chembiol.2017.01.008
32. Cummings M, Breitling R, Takano E (2014) Steps towards the synthetic biology of polyketide biosynthesis. FEMS Microbiol Lett 351(2):116–125. https://doi.org/10.1111/1574-6968.12365
33. Smith S, Tsai S-C (2007) The type I fatty acid and polyketide synthases: a tale of two megasynthases. Nat Prod Rep 24(5):1041–1072. https://doi.org/10.1039/b603600g
34. Robbins T, Kapilivsky J, Cane DE, Khosla C (2016) Roles of conserved active site residues in the ketosynthase domain of an assembly line polyketide synthase. Biochemistry 55(32):4476–4484. https://doi.org/10.1021/acs.biochem.6b00639
35. Olsen JG, Kadziola A, von Wettstein-Knowles P, Siggaard-Andersen M, Larsen S (2001) Structures of beta-ketoacyl-acyl carrier protein synthase I complexed with fatty acids elucidate its catalytic machinery. Structure 9(3):233–243. https://doi.org/10.1016/s0969-2126(01)00583-4
36. Witkowski A, Joshi AK, Smith S (2002) Mechanism of the beta-ketoacyl synthase reaction catalyzed by the animal fatty acid synthase. Biochemistry 41(35):10877–10887. https://doi.org/10.1021/bi0259047
37. Lowry B, Li XY, Robbins T, Cane DE, Khosla C (2016) A turnstile mechanism for the controlled growth of biosynthetic intermediates on assembly line polyketide synthases. Acs Central Sci 2(1):14–20. https://doi.org/10.1021/acscentsci.5b00321
38. Broadhurst RW, Nietlispach D, Wheatcroft MP, Leadlay PF, Weissman KJ (2003) The structure of docking domains in modular polyketide synthases. Chem Biol 10(8):723–731. https://doi.org/10.1016/S1074-5521(03)00156-X
39. Dodge GJ, Maloney FP, Smith JL (2018) Protein-protein Interactions in "cis-AT" polyketide synthases. Nat Prod Rep 35(10):1082–1096. https://doi.org/10.1039/c8np00058a
40. Kosol S, Jenner M, Lewandowski JR, Challis GL (2018) Protein-protein interactions in trans-AT polyketide synthases. Nat Prod Rep 35(10):1097–1109. https://doi.org/10.1039/c8np00066b
41. Yu DY, Xu FC, Zeng J, Zhan JX (2012) Type III polyketide synthases in natural product biosynthesis. IUBMB Life 64(4):285–295. https://doi.org/10.1002/iub.1005
42. Schumann J, Hertweck C (2006) Advances in cloning, functional analysis and heterologous expression of fungal polyketide synthase genes. J Biotechnol 124(4):690–703. https://doi.org/10.1016/j.jbiotec.2006.03.046
43. Chen HT, Du LC (2016) Iterative polyketide biosynthesis by modular polyketide synthases in bacteria. Appl Microbiol Biotechnol 100(2):541–557. https://doi.org/10.1007/s00253-015-7093-0
44. Kroken S, Glass NL, Taylor JW, Yoder OC, Turgeon BG (2003) Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. Proc Natl Acad Sci USA 100(26):15670–15675. https://doi.org/10.1073/pnas.2532165100
45. Maier T, Leibundgut M, Ban N (2008) The crystal structure of a mammalian fatty acid synthase. Science 321(5894):1315–1322. https://doi.org/10.1126/science.1161269
46. Hardwicke MA, Rendina AR, Williams SP, Moore ML, Wang L, Krueger JA, Plant RN, Totoritis RD, Zhang G, Briand J, Burkhart WA, Brown KK, Parrish CA (2014) A human fatty acid synthase inhibitor binds beta-ketoacyl reductase in the keto-substrate site. Nat Chem Biol 10(9):774–779. https://doi.org/10.1038/Nchembio.1603
47. Roberts DM, Bartel C, Scott A, Ivison D, Simpson TJ, Cox RJ (2017) Substrate selectivity of an isolated enoyl reductase catalytic domain from an iterative highly reducing fungal polyketide synthase reveals key components of programming. Chem Sci 8(2):1116–1126. https://doi.org/10.1039/c6sc03496a
48. Fisch KM, Bakeer W, Yakasai AA, Song ZS, Pedrick J, Wasil Z, Bailey AM, Lazarus CM, Simpson TJ, Cox RJ (2011) Rational domain swaps decipher programming in fungal highly reducing polyketide synthases and resurrect an extinct metabolite. J Am Chem Soc 133(41):16635–16641. https://doi.org/10.1021/ja206914q
49. Cacho RA, Thuss J, Xu W, Sanichar R, Gao ZZ, Nguyen A, Vederas JC, Tang Y (2015) Understanding programming of fungal iterative polyketide synthases: the biochemical basis for regioselectivity by the methyltransferase domain in the lovastatin megasynthase. J Am Chem Soc 137(50):15688–15691. https://doi.org/10.1021/jacs.5b11814
50. Viegas MF, Neves RPP, Ramos MJ, Fernandes PA (2018) Modeling of human fatty acid synthase and in silico docking of acyl carrier protein domain and its partner catalytic domains. J Phys Chem B 122(1):77–85. https://doi.org/10.1021/acs.jpcb.7b09645
51. Crawford JM, Thomas PM, Scheerer JR, Vagstad AL, Kelleher NL, Townsend CA (2008) Deconstruction of

iterative multidomain polyketide synthase function. Science 320(5873):243–246. https://doi.org/10.1126/science.1154711
52. Newman AG, Vagstad AL, Storm PA, Townsend CA (2014) Systematic domain swaps of iterative, nonreducing polyketide synthases provide a mechanistic understanding and rationale for catalytic reprogramming. J Am Chem Soc 136(20):7348–7362. https://doi.org/10.1021/ja5007299
53. Crawford JM, Dancy BCR, Hill EA, Udwary DW, Townsend CA (2006) Identification of a starter unit acyl-carrier protein transacylase domain in an iterative type I polyketide synthase. Proc Natl Acad Sci USA 103(45):16728–16733. https://doi.org/10.1073/pnas.0604112103
54. Crawford JM, Korman TP, Labonte JW, Vagstad AL, Hill EA, Kamari-Bidkorpeh O, Tsai SC, Townsend CA (2009) Structural basis for biosynthetic programming of fungal aromatic polyketide cyclization. Nature 461(7267):1139-U1243. https://doi.org/10.1038/nature08475
55. Fujii I, Watanabe A, Sankawa U, Ebizuka Y (2001) Identification of Claisen cyclase domain in fungal polyketide synthase WA, a naphthopyrone synthase of Aspergillus nidulans. Chem Biol 8(2):189–197. https://doi.org/10.1016/S1074-5521(00)90068-1
56. Anand S, Mohanty D (2012) Inter-domain movements in polyketide synthases: a molecular dynamics study. Mol Biosyst 8(4):1157–1171. https://doi.org/10.1039/c2mb05425f
57. Nivina A, Yuet KP, Hsu J, Khosla C (2019) Evolution and diversity of assembly-line polyketide synthases. Chem Rev 119(24):12524–12547. https://doi.org/10.1021/acs.chemrev.9b00525
58. Anand S, Mohanty D (2012) Modeling holo-ACP:DH and holo-ACP:KR complexes of modular polyketide synthases: a docking and molecular dynamics study. BMC Struct Biol 12(1):10. https://doi.org/10.1186/1472-6807-12-10
59. Keatinge-Clay AT, Stroud RM (2006) The structure of a ketoreductase determines the organization of the beta-carbon processing enzymes of modular polyketide synthases. Structure 14(4):737–748. https://doi.org/10.1016/j.str.2006.01.009
60. Leesong M, Henderson BS, Gillig JR, Schwab JM, Smith JL (1996) Structure of a dehydratase–isomerase from the bacterial pathway for biosynthesis of unsaturated fatty acids: two catalytic activities in one active site. Structure 4(3):253–264. https://doi.org/10.1016/s0969-2126(96)00030-5
61. Mugnai ML, Shi Y, Keatinge-Clay AT, Elber R (2015) Molecular dynamics studies of modular polyketide synthase ketoreductase stereospecificity. Biochemistry 54(14):2346–2359. https://doi.org/10.1021/bi501401g
62. Liu L, Yu Q, Zhang H, Tao W, Wang R, Bai L, Zhao Y-L, Shi T (2021) Theoretical study on substrate recognition and catalytic mechanisms of gephyronic acid dehydratase DH1. Catal Sci Technol 11(6):2155–2166. https://doi.org/10.1039/D0CY01776K
63. Shi T, Liu LX, Tao WT, Luo SG, Fan SB, Wang XL, Bai LQ, Zhao YL (2018) Theoretical studies on the catalytic mechanism and substrate diversity for macrocyclization of pikromycin thioesterase. ACS Catal 8(5):4323–4332. https://doi.org/10.1021/acscatal.8b01156
64. Chen XP, Shi T, Wang XL, Wang JT, Chen QH, Bai LQ, Zhao YL (2016) Theoretical studies on the mechanism of thioesterase-catalyzed macrocyclization in erythromycin biosynthesis. ACS Catal 6(7):4369–4378. https://doi.org/10.1021/acscatal.6b01154
65. Liu L, Tao WT, Bai LQ, Kim ES, Zhao YL, Shi T (2019) Why does tautomycetin thioesterase prefer hydrolysis to macrocyclization? Theoretical study on its catalytic mechanism. Catal Sci Technol 9(22):6391–6403. https://doi.org/10.1039/c9cy01355e
66. Burson KK, Khosla C (2000) Dissecting the w. Tetrahedron 56(48):9401–9408. https://doi.org/10.1016/S0040-4020(00)00824-3
67. Tang Y, Tsai SC, Khosla C (2003) Polyketide chain length control by chain length factor. J Am Chem Soc 125(42):12708–12709. https://doi.org/10.1021/ja0378759
68. Nicholson TP, Winfield C, Westcott J, Crosby J, Simpson TJ, Cox RJ (2003) First in vitro directed biosynthesis of new compounds by a minimal type II polyketide synthase: evidence for the mechanism of chain length determination. Chem Commun 6:686–687. https://doi.org/10.1039/b300847a
69. Shakya G, Rivera H, Lee DJ, Jaremko MJ, La Clair JJ, Fox DT, Haushalter RW, Schaub AJ, Bruegger J, Barajas JF, White AR, Kaur P, Gwozdziowski ER, Wong F, Tsai SC, Burkart MD (2014) Modeling linear and cyclic PKS intermediates through atom replacement. J Am Chem Soc 136(48):16792–16799. https://doi.org/10.1021/ja5064857
70. Brauer A, Zhou QQ, Grammbitter GLC, Schmalhofer M, Ruhl M, Kaila VRI, Bode HB, Groll M (2020) Structural snapshots of the minimal PKS system responsible for octaketide biosynthesis. Nat Chem. https://doi.org/10.1038/s41557-020-0491-7
71. Ellis BD, Milligan JC, White AR, Duong V, Altman PX, Mohammed LY, Crump MP, Crosby J, Luo R, Vanderwal CD, Tsai SC (2018) An oxetane-based polyketide surrogate to probe substrate binding in a polyketide synthase. J Am Chem Soc 140(15):4961–4964. https://doi.org/10.1021/jacs.7b11793
72. Pohle S, Appelt C, Roux M, Fiedler HP, Sussmuth RD (2011) Biosynthetic gene cluster of the non-ribosomally synthesized cyclodepsipeptide skyllamycin: deciphering unprecedented ways of unusual hydroxylation reactions. J Am Chem Soc 133(16):6194–6205. https://doi.org/10.1021/ja108971p
73. Bilyk O, Brotz E, Tokovenko B, Bechthold A, Paululat T, Luzhetskyy A (2016) New simocyclinones: surprising evolutionary and biosynthetic insights. ACS Chem Biol 11(1):241–250. https://doi.org/10.1021/acschembio.5b00669
74. Du D, Katsuyama Y, Shin-Ya K, Ohnishi Y (2018) Reconstitution of a type II polyketide synthase that catalyzes polyene formation. Angew Chem Int Edit 57(7):1954–1957. https://doi.org/10.1002/anie.201709636
75. Mcdaniel R, Ebertkhosla S, Fu H, Hopwood DA, Khosla C (1994) Engineered biosynthesis of novel polyketides—influence of a downstream enzyme on the catalytic specificity of a minimal aromatic polyketide synthase. Proc Natl Acad Sci USA 91(24):11542–11546. https://doi.org/10.1073/pnas.91.24.11542
76. Hertweck C, Xiang LK, Kalaitzis JA, Cheng Q, Palzer M, Moore BS (2004) Context-dependent behavior of the enterocin iterative polyketide synthase: a new model for ketoreduction. Chem Biol 11(4):461–468. https://doi.org/10.1016/j.chembiol.2004.03.018
77. Javidpour P, Bruegger J, Srithahan S, Korman TP, Crump MP, Crosby J, Burkart MD, Tsai SC (2013) The determinants of activity and specificity in actinorhodin type II polyketide ketoreductase. Chem Biol 20(10):1225–1234. https://doi.org/10.1016/j.chembiol.2013.07.016
78. Valentic TR, Jackson DR, Brady SF, Tsai SC (2016) Comprehensive analysis of a novel ketoreductase for pentangular polyphenol biosynthesis. ACS Chem Biol 11(12):3421–3430. https://doi.org/10.1021/acschembio.6b00658
79. Serapian SA, van der Kamp MW (2019) Unpicking the cause of stereoselectivity in actinorhodin ketoreductase variants with atomistic simulations. ACS Catal 9(3):2381–2394. https://doi.org/10.1021/acscatal.8b04846
80. Mcdaniel R, Ebertkhosla S, Hopwood DA, Khosla C (1994) Engineered biosynthesis of novel polyketides—Act(VII) and Act(IV) genes encode aromatase and cyclase enzymes Respectively. J Am Chem Soc 116(24):10855–10859. https://doi.org/10.1021/ja00103a001
81. Shimizu Y, Ogata H, Goto S (2017) Type III polyketide synthases: functional classification and phylogenomics. ChemBioChem 18(1):50–65. https://doi.org/10.1002/cbic.201600522

82. Liu B, Raeth T, Beuerle T, Beerhues L (2007) Biphenyl synthase, a novel type III polyketide synthase. Planta 225(6):1495–1503. https://doi.org/10.1007/s00425-006-0435-5
83. Ferrer JL, Jez JM, Bowman ME, Dixon RA, Noel JP (1999) Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis. Nat Struct Biol 6(8):775–784. https://doi.org/10.1038/11553
84. Jez JM, Bowman ME, Noel JP (2002) Expanding the biosynthetic repertoire of plant type III polyketide synthases by altering starter molecule specificity. Proc Natl Acad Sci USA 99(8):5319–5324. https://doi.org/10.1073/pnas.082590499
85. Song LJ, Barona-Gomez F, Corre C, Xiang LK, Udwary DW, Austin MB, Noel JP, Moore BS, Challis GL (2006) Type III polyketide synthase beta-ketoacyl-ACP starter unit and ethylmalonyl-CoA extender unit selectivity discovered by streptomyces coelicolor genome mining. J Am Chem Soc 128(46):14754–14755. https://doi.org/10.1021/ja065247w
86. Funabashi M, Funa N, Horinouchi S (2008) Phenolic lipids synthesized by type III polyketide synthase confer penicillin resistance on Streptomyces griseus. J Biol Chem 283(20):13983–13991. https://doi.org/10.1074/jbc.M710461200
87. Austin MB, Noel AJP (2003) The chalcone synthase superfamily of type III polyketide synthases. Nat Prod Rep 20(1):79–110. https://doi.org/10.1039/b100917f
88. Rittner A, Paithankar KS, Himmler A, Grininger M (2020) Type I fatty acid synthase trapped in the octanoyl-bound state. Protein Sci 29(2):589–605. https://doi.org/10.1002/pro.3797
89. Healy EF, Cervantes L, Nabona B, Williams J (2018) A unified mechanism for plant polyketide biosynthesis derived from in silico modeling. Biochem Biophys Res Commun 497(4):1123–1128. https://doi.org/10.1016/j.bbrc.2018.02.190
90. Yamaguchi T, Kurosaki F, Suh DY, Sankawa U, Nishioka M, Akiyama T, Shibuya M, Ebizuka Y (1999) Cross-reaction of chalcone synthase and stilbene synthase overexpressed in Escherichia coli. FEBS Lett 460(3):457–461. https://doi.org/10.1016/S0014-5793(99)01403-9
91. Samappito S, Page JE, Schmidt J, De-Eknamkul W, Kutchan TM (2003) Aromatic and pyrone polyketides synthesized by a stilbene synthase from Rheum tataricum. Phytochemistry 62(3):313–323. https://doi.org/10.1016/S0031-9422(02)00545-9
92. Anand S, Prasad MVR, Yadav G, Kumar N, Shehara J, Ansari MZ, Mohanty D (2010) SBSPKS: structure based sequence analysis of polyketide synthases. Nucleic Acids Res 38:W487–W496. https://doi.org/10.1093/nar/gkq340
93. Mallika V, Sivakumar K, Jaichand S, Soniya E (2010) Kernel based machine learning algorithm for the efficient prediction of type III polyketide synthase family of proteins. J Integr Bioinform 7(1):47–54. https://doi.org/10.2390/biecoll-jib-2010-143
94. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Muller R, Wohlleben W, Breitling R, Takano E, Medema MH (2015) antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res 43:W237–W243. https://doi.org/10.1093/nar/gkv437
95. Menzella HG, Reid R, Carney JR, Chandran SS, Reisinger SJ, Patel KG, Hopwood DA, Santi DV (2005) Combinatorial polyketide biosynthesis by de novo design and rearrangement of modular polyketide synthase genes. Nat Biotechnol 23(9):1171–1176. https://doi.org/10.1038/nbt1128
96. Miyazawa T, Hirsch M, Zhang ZC, Keatinge-Clay AT (2020) An in vitro platform for engineering and harnessing modular polyketide synthases. Nat Commun 11(1):80. https://doi.org/10.1038/s41467-019-13811-0
97. Koehn FE (2012) Biosynthetic medicinal chemistry of natural product drugs. Medchemcomm 3(8):854–865. https://doi.org/10.1039/c2md00316c
98. Wlodek A, Kendrew SG, Coates NJ, Hold A, Pogwizd J, Rudder S, Sheehan LS, Higginbotham SJ, Stanley-Smith AE, Warneck T, Nur-E-Alam M, Radzom M, Martin CJ, Overvoorde L, Samborskyy M, Alt S, Heine D, Carter GT, Graziani EI, Koehn FE, McDonald L, Alanine A, Sarmiento RMR, Chao SK, Ratni H, Steward L, Norville IH, Sarkar-Tyson M, Moss SJ, Leadlay PF, Wilkinson B, Gregory MA (2017) Diversity oriented biosynthesis via accelerated evolution of modular gene clusters. Nat Commun 8(1):1206. https://doi.org/10.1038/s41467-017-01344-3
99. Hranueli D, Cullum J, Basrak B, Goldstein P, Long PF (2005) Plasticity of the streptomyces genome-evolution and engineering of new antibiotics. Curr Med Chem 12(14):1697–1704. https://doi.org/10.2174/0929867054367176
100. Weissman KJ, Leadlay PF (2005) Combinatorial biosynthesis of reduced polyketides. Nat Rev Microbiol 3(12):925–936. https://doi.org/10.1038/nrmicro1287
101. Zucko J, Cullum J, Hranueli D, Long PF (2011) Evolutionary dynamics of modular polyketide synthases, with implications for protein design and engineering. J Antibiot 64(1):89–92. https://doi.org/10.1038/ja.2010.141
102. Gokhale RS, Tsuji SY, Cane DE, Khosla C (1999) Dissecting and exploiting intermodular communication in polyketide synthases. Science 284(5413):482–485. https://doi.org/10.1126/science.284.5413.482
103. Ranganathan A, Timoney M, Bycroft M, Cortes J, Thomas IP, Wilkinson B, Kellenberger L, Hanefeld U, Galloway IS, Staunton J, Leadlay PF (1999) Knowledge-based design of bimodular and trimodular polyketide synthases based on domain and module swaps: a route to simple statin analogues. Chem Biol 6(10):731–741. https://doi.org/10.1016/S1074-5521(00)80020-4
104. Tang L, Fu H, McDaniel R (2000) Formation of functional heterologous complexes using subunits from the picromycin, erythromycin and oleandomycin polyketide synthases. Chem Biol 7(2):77–84. https://doi.org/10.1016/S1074-5521(00)00073-9
105. Lowry B, Robbins T, Weng CH, O'Brien RV, Cane DE, Khosla C (2013) In vitro reconstitution and analysis of the 6-deoxyerythronolide B synthase. J Am Chem Soc 135(45):16809–16812. https://doi.org/10.1021/ja409048k
106. Zhang LH, Hashimoto T, Qin B, Hashimoto J, Kozone I, Kawahara T, Okada M, Awakawa T, Ito T, Asakawa Y, Ueki M, Takahashi S, Osada H, Wakimoto T, Ikeda H, Shin-ya K, Abe I (2017) Characterization of giant modular pkss provides insight into genetic mechanism for structural diversification of aminopolyol polyketides. Angew Chem Int Edit 56(7):1740–1745. https://doi.org/10.1002/anie.201611371
107. Barajas JF, Blake-Hedges JM, Bailey CB, Curran S, Keasling JD (2017) Engineered polyketides: synergy between protein and host level engineering. Synth Syst Biotechno 2(3):147–166. https://doi.org/10.1016/j.synbio.2017.08.005
108. Hans M, Hornung A, Dziarnowski A, Cane DE, Khosla C (2003) Mechanistic analysis of acyl transferase domain exchange in polyketide synthase modules. J Am Chem Soc 125(18):5366–5374. https://doi.org/10.1021/ja029539i
109. Yuzawa S, Deng K, Wang G, Baidoo EEK, Northen TR, Adams PD, Katz L, Keasling JD (2017) Comprehensive in vitro analysis of acyltransferase domain exchanges in modular polyketide synthases and its application for short-chain ketone production. Acs Synth Biol 6(1):139–147. https://doi.org/10.1021/acssynbio.6b00176
110. Annaval T, Paris C, Leadlay PF, Jacob C, Weissman KJ (2015) Evaluating ketoreductase exchanges as a means of rationally altering polyketide stereochemistry. ChemBioChem 16(9):1357–1364. https://doi.org/10.1002/cbic.201500113
111. Drufva EE, Hix EG, Bailey CB (2020) Site directed mutagenesis as a precision tool to enable synthetic biology with engineered

modular polyketide synthases. Synth Syst Biotechno 5(2):62–80. https://doi.org/10.1016/j.synbio.2020.04.001
112. Morita H, Wong CP, Abe I (2019) How structural subtleties lead to molecular diversity for the type III polyketide synthases. J Biol Chem 294(41):15121–15136. https://doi.org/10.1074/jbc.REV119.006129
113. Keatinge-Clay AT (2016) Stereocontrol within polyketide assembly lines. Nat Prod Rep 33(2):141–149. https://doi.org/10.1039/c5np00092k
114. Murphy AC, Hong H, Vance S, Broadhurst RW, Leadlay PF (2016) Broadening substrate specificity of a chain-extending ketosynthase through a single active-site mutation. Chem Commun 52(54):8373–8376. https://doi.org/10.1039/c6cc03501a
115. Dunn BJ, Watts KR, Robbins T, Cane DE, Khosla C (2014) Comparative analysis of the substrate specificity of trans- versus cis-acyltransferases of assembly line polyketide synthases. Biochemistry 53(23):3796–3806. https://doi.org/10.1021/bi5004316
116. Musiol-Kroll EM, Zubeil F, Schafhauser T, Hartner T, Kulik A, McArthur J, Koryakina I, Wohlleben W, Grond S, Williams GJ, Lee SY, Weber T (2017) Polyketide bioderivatization using the promiscuous acyltransferase KirCII. Acs Synth Biol 6(3):421–427. https://doi.org/10.1021/acssynbio.6b00341
117. Dunn BJ, Khosla C (2013) Engineering the acyltransferase substrate specificity of assembly line polyketide synthases. J R Soc Interface 10(85):20130297. https://doi.org/10.1098/rsif.2013.0297
118. Kalkreuter E, Bingham KS, Keeler AM, Lowell AN, Schmidt JJ, Sherman DH, Williams GJ (2021) Computationally-guided exchange of substrate selectivity motifs in a modular polyketide synthase acyltransferase. Nat Commun. https://doi.org/10.1038/s41467-021-22497-2
119. Minotti G, Menna P, Salvatorelli E, Cairo G, Gianni L (2004) Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity. Pharmacol Rev 56(2):185–229. https://doi.org/10.1124/pr.56.2.6
120. Griffin MO, Fricovsky E, Ceballos G, Villarreal F (2010) Tetracyclines: a pleitropic family of compounds with promising therapeutic properties. Review of the literature. Am J Physiol Cell 299(3):C539–C548. https://doi.org/10.1152/ajpcell.00047.2010
121. Wang J, Zhang RH, Chen X, Sun XX, Yan YJ, Shen XL, Yuan QP (2020) Biosynthesis of aromatic polyketides in microorganisms using type II polyketide synthases. Microb Cell Fact. https://doi.org/10.1186/s12934-020-01367-4
122. Waldman AJ, Balskus EP (2014) Lomaiviticin biosynthesis employs a new strategy for starter unit generation. Org Lett 16(2):640–643. https://doi.org/10.1021/ol403714g
123. Pokhrel AR, Chaudhary AK, Nguyen HT, Dhakal D, Le TT, Shrestha A, Liou K, Sohng JK (2016) Overexpression of a pathway specific negative regulator enhances production of daunorubicin in bldA deficient Streptomyces peucetius ATCC 27952. Microbiol Res 192:96–102. https://doi.org/10.1016/j.micres.2016.06.009
124. Wang XR, Tian XR, Wu YJ, Shen XF, Yang SB, Chen SX (2018) Enhanced doxorubicin production by Streptomyces peucetius using a combination of classical strain mutation and medium optimization. Prep Biochem Biotechnol 48(6):514–521. https://doi.org/10.1080/10826068.2018.1466156
125. Thomas MG, Burkart MD, Walsh CT (2002) Conversion of L-proline to pyrrolyl-2-carboxyl-S-PCP during undecylprodigiosin and pyoluteorin biosynthesis. Chem Biol 9(2):171–184. https://doi.org/10.1016/S1074-5521(02)00100-X
126. Gan ML, Liu B, Tan Y, Wang Q, Zhou HX, He HW, Ping YH, Yang ZY, Wang YG, Xiao CL (2015) Saccharothrixones A-D, tetracenomycin-type polyketides from the marine-derived actinomycete saccharothrix sp 10–10. J Nat Prod 78(9):2260–2265. https://doi.org/10.1021/acs.jnatprod.5b00577
127. Jiang L, Pu H, Xiang JX, Su M, Yan XH, Yang D, Zhu XC, Shen B, Duan YW, Huang Y (2018) Huanglongmycin A-C, cytotoxic polyketides biosynthesized by a putative type II polyketide synthase from streptomyces sp CB09001. Front Chem. https://doi.org/10.3389/fchem.2018.00254
128. Liu Q, Wu KY, Cheng YB, Lu L, Xiao ET, Zhang YC, Deng ZX, Liu TG (2015) Engineering an iterative polyketide pathway in Escherichia coli results in single-form alkene and alkane overproduction. Metab Eng 28:82–90. https://doi.org/10.1016/j.ymben.2014.12.004
129. Hagen A, Poust S, de Rond T, Fortman JL, Katz L, Petzold CJ, Keasling JD (2016) Engineering a polyketide synthase for in vitro production of adipic acid. Acs Synth Biol 5(1):21–27. https://doi.org/10.1021/acssynbio.5b00153
130. Ray L, Moore BS (2016) Recent advances in the biosynthesis of unusual polyketide synthase substrates. Nat Prod Rep 33(2):150–161. https://doi.org/10.1039/c5np00112a
131. Cai WL, Zhang WJ (2018) Engineering modular polyketide synthases for production of biofuels and industrial chemicals. Curr Opin Biotechnol 50:32–38. https://doi.org/10.1016/j.copbio.2017.08.017
132. Eng CH, Backman TWH, Bailey CB, Magnan C, Martin HG, Katz L, Baldi P, Keasling JD (2018) ClusterCAD: a computational platform for type I modular polyketide synthase design. Nucleic Acids Res 46(D1):D509–D515. https://doi.org/10.1093/nar/gkx893
133. Zargar A, Lal R, Valencia L, Wang J, Backman TWH, Cruz-Morales P, Kothari A, Werts M, Wong AR, Bailey CB, Loubat A, Liu YZ, Chen Y, Chang S, Benites VT, Hernandez AC, Barajas JF, Thompson MG, Barcelos C, Anayah R, Martin HG, Mukhopadhyay A, Petzold CJ, Baidoo EEK, Katz L, Keasling JD (2020) Chemoinformatic-guided engineering of polyketide synthases. J Am Chem Soc 142(22):9896–9901. https://doi.org/10.1021/jacs.0c02549
134. Sundermann U, Bravo-Rodriguez K, Klopries S, Kushnir S, Gomez H, Sanchez-Garcia E, Schulz F (2013) Enzyme-directed mutasynthesis: a combined experimental and theoretical approach to substrate recognition of a polyketide synthase. ACS Chem Biol 8(2):443–450. https://doi.org/10.1021/cb300505w
135. Radom F, Pluckthun A, Paci E (2018) Assessment of ab initio models of protein complexes by molecular dynamics. PLoS Comput Biol 14(6):1006182. https://doi.org/10.1371/journal.pcbi.1006182
136. Pan AC, Jacobson D, Yatsenko K, Sritharan D, Weinreich TM, Shaw DE (2019) Atomic-level characterization of protein-protein association. Proc Natl Acad Sci USA 116(10):4244–4249. https://doi.org/10.1073/pnas.1815431116
137. Zhang F, Shi T, Ji HN, Ali I, Huang SX, Deng ZX, Min Q, Bai LQ, Zhao YL, Zheng JT (2019) Structural insights into the substrate specificity of acyltransferases from salinomycin polyketide synthase. Biochemistry 58(27):2978–2986. https://doi.org/10.1021/acs.biochem.9b00305
138. Korendovych IV (2018) Rational and semirational protein design. In: Bornscheuer UT, Höhne M (eds) Protein engineering: methods and protocols. Springer, New York
139. Rigoldi F, Donini S, Redaelli A, Parisini E, Gautieri A (2018) Review: Engineering of thermostable enzymes for industrial applications. Apl Bioeng. DOI 10(1063/1):4997367

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

CHAPTER 4. Study of Glucuronoyl Esterase

In chapter 1, an extensive introduction and contextualization were given to the importance of green chemistry to devise strategies to curb climate change. Biofuels, alongside more efficient processing and use of waste biomass, are important contributors to a CO_2-neutral society.

Massive bioethanol production from lignocellulosic biomass will contribute to this goal. Still, nowadays, it is hampered by the complex process of lignocellulosic material's breakdown for fermentable extraction sugars such as cellulose. The breakdown of lignocellulosic material through green processes demands a plethora of biocatalysts. Extremophilic enzymes possess a number of features that differentiate them from their mesophilic counterparts. These features are summarized in the introductory mini review paper "Thermophilic Enzymes: The Key for Greener Industrial Processes" available at the beginning of this chapter and explain the advantages of using extremophilic enzymes as biocatalysts.

In the context of lignocellulosic biomass processing, LCC complexes are responsible for the integrity of the plant cell wall, but they also pose a major obstacle due to biomass recalcitrance. This impedes successful utilization of lignocellulosic biomass. Therefore, the interest in glucuronoyl esterases has intensified in the past decade due to their potential in degrading LCC complexes. Over the course of this thesis, fungal *Thermothelomyces thermophila* glucuronoyl esterase (*Tt*GE) was studied. This enzyme is thermo-resistant and a potential candidate for biocatalysis in the context of LCC processing. Using QM/MM ONIOM methods, we were able to determine the catalytic mechanism of *Tt*GE with an atomistic detail, and determine the limiting step and the energetic barrier of the catalyzed reaction. We also performed *in silico* mutagenesis assays that suggested mutations that can potentially decrease the energetic expenditure, increasing its enzymatic efficiency.

Altogether, our studies aim to contribute to the possibility of using these biocatalysts in harsh industrial conditions. The shift to biocatalysts is a step closer to more environmental-friendly industrial processing of waste lignocellulosic biomass.

4.1. Thermophilic Enzymes: The Key for Greener Industrial Processes

Matilde F. Viegas[a], Maria J. Ramos[a], and Pedro A. Fernandes[a]

[a] LAQV@REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

---

Viegas, M., Ramos, M. J., & Alexandrino Fernandes, P. (2021). ***Thermophilic enzymes: The key for greener industrial processes.*** In U. Porto Journal of Engineering, 7(3), 13–23. https://doi.org/10.24840/2183-6493_007.003_0002

4.1.1. Abstract

Substantial improvements in the industrial production of goods led to a widespread feeling of unlimited access to food, commodities, and energy. As greener alternatives for industrial processes are in demand, scientists have turned to enzymes, looking for apt biocatalysts. Focusing on extremophiles, this mini review draws a comparison between thermophiles and their mesophilic counterparts, exploring what features are instrumental to their thermostability. A higher number of ion-pairs, hydrophobicity of buried side chains, compact tertiary structure cores, and a complex network of hydrogen bonds are the four main characteristics responsible for the robustness of thermophilic enzymes.

Keywords: Enzymes, Mesophiles, Thermophiles, Extremophiles, Protein Engineering.

U.Porto Journal of Engineering, 7:3 (2021) 13-23
ISSN 2183-6493
DOI: 10.24840/2183-6493_007.003_0002

Received: 13 May, 2020
Accepted: 4 June, 2020
Published: 30 April, 2021

Thermophilic Enzymes: The Key for Greener Industrial Processes

Matilde F. Viegas[1], Maria J. Ramos[2], Pedro A. Fernandes[3]
[1]LAQV@REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal (matilde.viegas@fc.up.pt) ORCID 0000-0002-4472-9691; [2]LAQV@REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal (mjramos@fc.up.pt) ORCID 0000-0002-7554-8324; [3]LAQV@REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal (pafernan@fc.up.pt) ORCID 0000-0003-2748-4722

Abstract
Substantial improvements in the industrial production of goods led to a widespread feeling of unlimited access to food, commodities, and energy. As greener alternatives for industrial processes are in demand, scientists have turned to enzymes, looking for apt biocatalysts. Focusing on extremophiles, this mini review draws a comparison between thermophiles and their mesophilic counterparts, exploring what features are instrumental to their thermostability. A higher number of ion-pairs, hydrophobicity of buried side chains, compact tertiary structure cores, and a complex network of hydrogen bonds are the four main characteristics responsible for the robustness of thermophilic enzymes.

Author Keywords. Enzymes, Mesophiles, Thermophiles, Extremophiles, Protein Engineering.

Type: Review Article
Open Access Peer Reviewed CC BY

1. Introduction

The world, as we know it, is tightly bound to the industrial advances of the twentieth century. The First and Second Industrial Revolutions, aligned with the use of fossil fuels in the nineteenth century, have prompted the world to evolve at a violent rate (Wrigley 2013). Substantial improvements in the industrial production of goods led to a widespread feeling of unlimited access to food, commodities, and energy. However, behind these are polluting industrial processes with major drawbacks to the environment, contributing to the rise of greenhouse gas emissions we have been witnessing during the last century (American Chemical Society 2016; United Nations 2008; 2015).

The concerns about the environment (Lewis 2016), and our future, have led scientists and industry sectors to invest both time and funds on research to find greener alternatives to industrial processes. Enzymes, the catalysts of a myriad of chemical reactions essential to life, offer high specificity and reaction rate enhancement, posing themselves as the scaffold for the search of greener solutions (Cipolatti et al. 2019).

Enzymes' best performance is typically achieved under physiological conditions as they are tremendously sensitive to any environmental fluctuations. However, a group of enzymes, called extremozymes, possesses the ability to perform the same biological tasks under severe settings. These enzymes belong to extremophile organisms that live in extreme conditions (high temperature, pressure, salt concentration, pH, among others) (Cavicchioli et al. 2011;

Deppe et al. 2005). Their adaptability makes them indisputably suited to harsh, industrial environments.

2. Enzymes, Nature's Manufacturers

Enzymes are responsible for accelerating (catalyzing) chemical reactions, lowering their activation energy, thus being crucial for life maintenance in living organisms (Berg, Tymoczko, and Stryer 2002). The use of enzymes in human-made products started 5000 years ago, in ancient Greece, when people ingeniously took advantage of rennet, a set of enzymes present in animal stomachs, to turn milk into cheese (Dekker 2019). Soon after, the use of microbial enzymes was extended to brewing and baking. In the past century, science has turned its interest to these macromolecules, intending to maximize its potential, and increasing its usage manifolds (Hao et al. 2011).

An archetype case of the use of enzymes is that of *Saccharomyces cerevisiae*. Despite the existence of thousands of microorganisms catalyzing glucose fermentation, *S. cerevisiae* has been extensively used in fermentation. It has since been studied, and genetic and metabolically engineered, garnering it its title of Baker's yeast (Deák 2003).

As is the case of *S. cerevisiae*, we can find a widespread use of enzymes in various industries. – food (Fernandes 2010), textile (Selvarajan, Veena, and Manoj Kumar 2018), animal feed (Walsh, Power, and Headon 1993; Cowieson, Hruby, and Pierson 2006; Cowieson, Masey-O´Neill, and Bedford 2013), paper (Singh et al. 2016; Srinivasan and Rele 1999; Bajpai 1999), biofuel production (d'Errico et al. 2016; Poppe et al. 2015; Alvira et al. 2010; Abin-Fuentes et al. 2013), pharmaceutical (Shrivastava, Shrivastava, and Singh 2018), and other markets – whose prime-matters are biomolecules that can be degraded, modified, or produced by enzymes. Enzymatic catalysis brings many advantages to the table, allowing for reduced energy input, lower costs, less polluting processes, and efficient product recovery at the end. Despite this evidence, only in the last two decades there has been a drastic increase in the use of biocatalysts (Choi, Han, and Kim 2015; Schmid et al. 2001). Currently, only about 20 enzymes are produced on an industrial scale, and 75% of these enzymes used in industrial settings are hydrolytic, degrading its substrate (Li et al. 2012). As technological advances allowed for its production and purification in industrial quantities, the global enzyme market size is proposed to reach USD 6.30 billion by 2022 (Selvarajan, Veena, and Manoj Kumar 2018). It is expectable that this industry will increase enormously in the fore coming decades as the scientific accomplishments start to raise interest from investors and industry.

In this review, we highlight and discuss structural characteristics of extremozymes, in particular, thermophilic enzymes, that allow for the maintenance of their function and structure in high or low temperatures.

3. Mesophilic vs. Extremophilic Enzymes

One may question why the sudden interest in enzymes when industrial processes have already been optimized using inorganic catalysts. Enzymes are a valuable and even preferable alternative: they are fast, specific, efficient, cost-effective, biodegradable, and do not produce as many by-products, resulting in a low environmental impact (Johnson 2013). However, the most commonly used enzymes come from fungi or mesophiles (organisms that grow best in temperatures between 20 °C and about 40°, typically 37ºC) (Rao, Boorgula, and Leitão 2011). Mesophilic enzymes exhibit their best performance under physiological conditions, still, in severe conditions such as the ones often encountered in industrial settings (extreme temperatures, pH and/or ionic strength) they are rendered useless. To overcome this, scientists have turned to extremophiles: organisms from the *Archea* and *Bacteria*, that thrive

in stark conditions such as the ones found in hot vents, deep sea, volcanoes, among others. The enzymes produced by these organisms, extremozymes, are heat- or cold-stable, and solvent-tolerant biocatalysts (Adrio and Demain 2014). Extremozymes possess useful native metabolic functions with potential for application in industrial and biotechnological processes. These include: cellulases (Ueda et al. 2010; Li et al. 2018), and other carbohydrate-degrading enzymes such as esterases (Charavgi et al. 2013), amylases (Agrawal, Dwevedi, and Kayastha 2019; Wu et al. 2018), xylanases (Basit et al. 2018; Sunna and Bergquist 2003), proteases (Jayakumar et al. 2012; Barzkar et al. 2018), pectinases (Kaur and Satyanarayana 2004), keratinases (Wu et al. 2017; Kornillowicz-Kowalska and Bohacz 2011), lipases (Schreck and Grunden 2014), peroxidases (Merlino et al. 2010) and phytases (Jatuwong et al. 2020; Berka et al. 1998).

In the past decades, the progress made in molecular genetics has allowed for extremozymes to be produced in recombinant mesophilic hosts, making it feasible to engineer and produce such peculiar enzymes for industrial and biotechnological use.

3.1. Extremozymes

Extremozymes are enzymes found in organisms deemed as extraordinary, given the conditions in which they can proliferate. When temperatures rise above 62ºC, mesophilic organisms perish, and only thermophilic prokaryotic organisms can survive. Above 80ºC, hyperthermophilic *Archea* organisms become dominating (Stetter 2011).

Regarding pH conditions, acidophiles grow optimally at pH values of 3.0 or below (Johnson and Schippers 2017). They are used in bioleaching, a process employing microorganisms such as bacteria or archaea for the extraction of valuable metals from low-grade ore, has long depended on mesophilic acidophiles such as *Acidithiobacillus, Ferroplasma, Leptospirillum* (Bosecker 1987; Jafari et al. 2019). Similarly, thermoacidophiles (growing at pH values around zero, and temperatures up to 65ºC) and thermoalkaliphiles (pH 6.0 - 10.0) exhibit clear applicability for industrial purposes, particularly hydrolases for the degradation of biomass for biofuel production (Sharma, Kawarabayasi, and Satyanarayana 2012; Zhao et al. 2018; Maruthamuthu and Van Elsas 2017; Zaparty and Siebers 2011). The group of extremozymes extends beyond pH and temperature adaptations; organisms thriving in high salinity environments are called halophiles (DasSarma and DasSarma 2015), while barophiles proliferate in extreme pressure environments (Horikoshi 1998).

Given the diverse conditions organisms live on planet Earth, investigating their adaptive mechanisms and structural features shine a light on how scientists can improve the enzyme's functionality and resilience for biotechnological applications.

3.2. Thermostability

The amino acid sequence of a protein dictates not only its function but also its three-dimensional structure. With this premise, genetically engineering a protein to improve its reaction rate, or alter its catalysis altogether, seems to be a straightforward process. To do so, scientists can use two distinct approaches, either rational design using site-directed mutagenesis based on prior knowledge or directed evolution, which simulates evolution (Balabanova et al. 2015; Ribeiro et al. 2019). However, when genetically tuning a protein, engineers will most often destabilize it, therefore facing the challenge of maintaining protein stability at the cost of desired functionality (Rubingh 1997).

Thermophiles carry with themselves envy-worth characteristics for biotechnological applications, such as temperature stability, solvent-tolerance, high affinity to the substrate. Instead of genetically altering a given mesophilic enzyme, tuning it for a specific industrial

application, using thermophiles as biocatalysts is an attractive and time-saving alternative (Elleuche et al. 2014).

From a general perspective, there are many different interactions and structural features contributing to thermal stability and denaturation resistance - a greater number of ionic interactions, a higher number of hydrophobic residues at the core of the protein, as well as a higher number of disulfide bridges (Yu and Huang 2014) and proline residues (Takano et al. 2009), resulting in greater packing, shorter surface loops, to name a few. Much like baking itself, it appears that thermal stability results from a careful balance of different ingredients, and not from a single one. Comparative studies have shown four significant determinants of structural stability weighing in the thermostability and resilience of thermophilic enzymes.

3.2.1. Ion pairs

Ion pairs are like- and oppositely charged groups, at a ≤4 Å distance, involved in the folding and stabilization of the tertiary structure of the protein and, consequently, playing an essential role in the protein's function. According to a study by Barlow and Thornton (1983), results show that from a selection of 38 globular proteins, one-third of the charged residues found in a protein are involved in ion-pairs, and 76% are responsible for stabilizing the tertiary structure of the protein.

Szilágyi and Závodszky (2000) established a close correlation between ion-pairs and thermophilic enzymes - extremely thermophilic proteins are characterized by stronger ion pairs than their mesophilic counterparts. These ion pairs were separated by less than 4 Å.

In 2013, another study focused on the ribosomal protein L7Ae from the hyperthermophilic *Aeropyrum pernix,* with a melting temperature of 110ºC, has detailed an extensive ion-pair network in the hyperthermophilic protein when compared to the human homolog (19 out of 24 ion pairs are involved in network *vs.* no ion-pairs in the human protein) (Bhuiya et al. 2013).

3.2.2. Hydrophobicity of buried side chains

The second most striking difference between thermophilic and mesophilic enzymes is the hydrophobic environment at their core. Gromiha et al. (2013) have shown that "80% of the thermophilic proteins examined were characterized by higher hydrophobicity than their mesophilic counterparts". Surprisingly, this was observed only at the core of the proteins, while the average hydrophobicity of the exterior residues was comparable between the two types of enzymes.

A comparative study on adenylate kinases (ADK) from the genus *Methanococcus* has shown that, despite the 68% amino acid sequence identity between thermophilic and mesophilic counterparts of methanococcal ADKs, their melting points were separated by 34ºC. This difference is attributable, in part, to a small number of non-conserved residues with hydrophobic side chains found in the core area, in close proximity with each other (Haney, Stees, and Konisky 1999). Recent studies, using different techniques (Vijayabaskar and Vishveshwara 2010; de Oliveira et al. 2018), support the same hypothesis.

This relevance of a highly hydrophobic core can be explained by the compaction level at the core of the protein. As the hydrophobic residues tightly adhere to one another, acting as an internal support structure counteracting destabilizing influences from the external environment, they provide a higher degree of thermal integrity.

3.2.3. Hydrogen bonds

Studies have shown that the presence and distribution of hydrogen bonds in thermophilic enzymes are distinctive when compared to mesophilic enzymes, as bonds established between exposed and buried residues tighten the protein into itself.

The thermos-stable firefly Luciferase has been widely used as a marker in bioimaging applications (Mezzanotte et al. 2017). In a study aiming to improve its stability by rational design, an amino acid substitution led to a disruption of hydrogen bonds in the mutant. This break of hydrogen bonds caused an increase in the protein's flexibility, therefore destabilizing it (Yu et al. 2015).

Bezsudnova et al. (2015) explored the influence of charged residues on the intramolecular hydrogen-bonding network of a polyextremophilic short-chain alcohol dehydrogenase. This protein is capable of enduring not only high temperature, but also high salinity, and the presence of organic solvents and denaturants. Their study showed that structural analysis showed an increase in the number of Charged-Neutral hydrogen bonds at the surface of the protein, compared to its thermophilic and mesophilic homologs. This indicates that an effective hydrogen bond crosslinking is instrumental in preventing instability when facing denaturation conditions.

3.2.4. Compact tertiary structure cores

Thermophilic proteins present a higher quantity of small amino acids rather than bulky ones (Tompa, Gromiha, and Saraboji 2016). Proline residues, characterized by its fixed dihedral angle, are widely distributed in transmembrane helices of integral membrane proteins (Brandl and Deber 1986; Sansom 1992). They possess a pyrrolidine ring that creates steric constraints in Pro-adjacent regions. Therefore, their role is to provide rigidity, leading to a decrease in the flexibility of the aminoacid chain (Li et al. 1996). Additionally, a higher number of disulfide bonds is found distributed across thermophilic proteins, which contribute to a tighter packing and higher resistance to unfolding (Ladenstein and Ren 2008; Jorda and Yeates 2011).

In a wide study performed with 400 (200 thermophilic and 200 mesophilic homologues) proteins, results demonstrate that factors contributing to core compactness are found across a large number of datasets of thermophilic enzymes. These include higher hydrophobicity, smaller volume nonpolar residues (Ala, Gly and Val), and smaller distribution of bulky polar residues (Panja, Bandopadhyay, and Maiti 2015).

4. Conclusions

Environmental concerns have brought awareness to the importance of optimization of industrial processes in order to reduce pollutants. For the past two decades, there has been an increase in research dedicated to biocatalysis: enzymes' intrinsic stereoselectivity, regioselectivity, and chemoselectivity is appealing for biotechnological applications and are, undeniably, greener solutions to chemical catalysts. Mesophilic enzymes present clear limits on the conditions they can operate on, however, extremophiles possess the characteristics to fit industrial catalysis. With technological advances in protein engineering, it is now possible to alter the enzyme's functionalities. However, these alterations require a fine-tuning of several factors. Extremophiles, with their increased stability and robustness, granter a more appropriate scaffold for protein engineering than mesophilic proteins. As demonstrated in the literature, extremophiles, and in particular thermophilic enzymes, can improve with, and support, the mutations intended to enhance an existing enzyme. In the future, as research progresses, a thorough understanding of the structural properties will enable the

development of equally efficient biocatalysts to replace polluting, costly chemical reactants in industrial processes.

References

Abin-Fuentes, A., S. Mohamed Mel, D. I. Wang, and K. L. Prather. 2013. "Exploring the mechanism of biocatalyst inhibition in microbial desulfurization". *Applied and Environmental Microbiology* 79, no. 24: 7807-17. https://doi.org/10.1128/AEM.02696-13.

Adrio, J. L., and A. L. Demain. 2014. "Microbial enzymes: Tools for biotechnological processes". *Biomolecules* 4, no. 1: 117-39. https://doi.org/10.3390/biom4010117.

Agrawal, D. C., A. Dwevedi, and A. M. Kayastha. 2019. "Biochemical and thermodynamic characterization of de novo synthesized beta-amylase from fenugreek". *International Journal of Biological Macromolecules* 130 (june): 786-97. https://doi.org/10.1016/j.ijbiomac.2019.02.162.

Alvira, P., E. Tomas-Pejo, M. Ballesteros, and M. J. Negro. 2010. "Pretreatment technologies for an efficient bioethanol production process based on enzymatic hydrolysis: A review". *Bioresource Technology* 101, no. 13: 4851-61. https://doi.org/10.1016/j.biortech.2009.11.093.

American Chemical Society. 2016. "Changes since the Industrial Revolution". https://www.acs.org/content/acs/en/climatescience/greenhousegases/industrialrevolution.html.

Bajpai, P. 1999. "Application of enzymes in the pulp and paper industry". *Biotechnology Progress* 15, no. 2: 147-57. https://doi.org/10.1021/bp990013k.

Balabanova, L., V. Golotin, A. Podvolotskaya, and V. Rasskazov. 2015. "Genetically modified proteins: Functional improvement and chimeragenesis". *Bioengineered* 6, no. 5: 262-74. https://doi.org/10.1080/21655979.2015.1075674.

Barlow, D. J., and J. M. Thornton. 1983. "Ion-pairs in proteins". *Journal of Molecular Biology* 168, no. 4 (august): 867-85. https://doi.org/10.1016/s0022-2836(83)80079-5.

Barzkar, N., A. Homaei, R. Hemmati, and S. Patel. 2018. "Thermostable marine microbial proteases for industrial applications: Scopes and risks". *Extremophiles* 22, no. 3 (may): 335-46. https://doi.org/10.1007/s00792-018-1009-8.

Basit, A., J. Liu, K. Rahim, W. Jiang, and H. Lou. 2018. "Thermophilic xylanases: From bench to bottle". *Critical Reviews in Biotechnology* 38, no. 7 (october): 989-1002. https://doi.org/10.1080/07388551.2018.1425662.

Berg, J. M., J. L. Tymoczko, and L. Stryer. 2002. *Biochemistry*. 5th ed. New York: W. H. Freeman.

Berka, R. M., M. W. Rey, K. M. Brown, T. Byun, and A. V. Klotz. 1998. "Molecular characterization and expression of a phytase gene from the thermophilic fungus Thermomyces lanuginosus". *Applied and Environmental Microbiology* 64, no. 11 (november): 4423-27. https://doi.org/10.1128/AEM.64.11.4423-4427.1998.

Bezsudnova, E. Y., T. E. Petrova, A. V. Popinako, M. Y. Antonov, T. N. Stekhanova, and V. O. Popov. 2015. "Intramolecular hydrogen bonding in the polyextremophilic short-chain dehydrogenase from the archaeon Thermococcus sibiricus and its close structural homologs". *Biochimie* 118: 82-9. https://doi.org/10.1016/j.biochi.2015.08.010.

Bhuiya, M. W., J. Suryadi, Z. Zhou, and B. A. Brown, 2nd. 2013. "Structure of the Aeropyrum pernix L7Ae multifunctional protein and insight into its extreme thermostability". *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* 69, no. 9 (september): 979-88. https://doi.org/10.1107/S1744309113021799.

Bosecker, K. 1987. "Microbial recycling of mineral waste products". *Acta Biotechnologica* 7, no. 6: 487-97. https://doi.org/10.1002/abio.370070602.

Brandl, C. J., and C. M. Deber. 1986. "Hypothesis about the function of membrane-buried proline residues in transport proteins". *Proceedings of the National Academy of Sciences of the United States of America* 83, no. 4: 917-21. https://doi.org/10.1073/pnas.83.4.917.

Cavicchioli, R., R. Amils, D. Wagner, and T. McGenity. 2011. "Life and applications of extremophiles". *Environmental Microbiology* 13, no. 8 (august): 1903-07. https://doi.org/10.1111/j.1462-2920.2011.02512.x.

Charavgi, M. D., M. Dimarogona, E. Topakas, P. Christakopoulos, and E. D. Chrysina. 2013. "The structure of a novel glucuronoyl esterase from Myceliophthora thermophila gives new insights into its role as a potential biocatalyst". *Acta Crystallographica Section D: Biological Crystallography* 69, no. 1 (january): 63-73. https://doi.org/10.1107/S0907444912042400.

Choi, J. M., S. S. Han, and H. S. Kim. 2015. "Industrial applications of enzyme biocatalysis: Current status and future aspects". *Biotechnology Advances* 33, no. 7 (november): 1443-54. https://doi.org/10.1016/j.biotechadv.2015.02.014.

Cipolatti, E. P., M. C. Cerqueira Pinto, R. O. Henriques, J. C. C. da Silva Pinto, A. M. de Castro, D. M. G. Freire, and E. A. Manoel. 2019. "Enzymes in green chemistry: The state of the art in chemical transformations". In *Advances in Enzyme Technology*, 137-51. Elsevier. https://doi.org/10.1016/B978-0-444-64114-4.00005-4.

Cowieson, A., H. Masey-O´Neill, and M. Bedford. 2013. "Enzymes beyond phytase in poultry nutrition". Poultry Research Foundation.

Cowieson, A. J., M. Hruby, and E. E. M. Pierson. 2006. "Evolving enzyme technology: Impact on commercial poultry nutrition". *Nutrition Research Reviews* 19, no. 1 (june): 90-103. https://doi.org/10.1079/Nrr2006121.

d'Errico, C., J. Borjesson, H. S. Ding, K. B. R. M. Krogh, N. Spodsberg, R. Madsen, and R. N. Monrad. 2016. "Improved biomass degradation using fungal glucuronoyl-esterases-hydrolysis of natural corn fiber substrate". *Journal of Biotechnology* 219 (february): 117-23. https://doi.org/10.1016/j.jbiotec.2015.12.024.

DasSarma, S., and P. DasSarma. 2015. "Halophiles and their enzymes: Negativity put to good use". *Current Opinion in Microbiology* 25 (june): 120-26. https://doi.org/10.1016/j.mib.2015.05.009.

de Oliveira, T. B., C. Gostincar, N. Gunde-Cimerman, and A. Rodrigues. 2018. "Genome mining for peptidases in heat-tolerant and mesophilic fungi and putative adaptations for thermostability". *Bmc Genomics* 19, no. 1 (february): Article number 152. https://doi.org/10.1186/s12864-018-4549-5.

Deák, T. 2003. "YEASTS". In *Encyclopedia of Food Sciences and Nutrition*, 6233-39. Oxford: Academic Press. https://doi.org/10.1016/B0-12-227055-X/01302-X.

Dekker, P. 2019. "Dairy enzymes". In *Industrial Enzyme Applications*, 143-66. Wiley-VCH Verlag GmbH & Co. KGaA. https://doi.org/10.1002/9783527813780.ch2_3.

Deppe, U., H. H. Richnow, W. Michaelis, and G. Antranikian. 2005. "Degradation of crude oil by an arctic microbial consortium". *Extremophiles* 9, no. 6 (december): 461-70. https://doi.org/10.1007/s00792-005-0463-2.

Elleuche, S., C. Schröder, K. Sahm, and G. Antranikian. 2014. "Extremozymes - Biocatalysts with unique properties from extremophilic microorganisms". *Current Opinion in Biotechnology* 29 (october): 116-23. https://doi.org/10.1016/j.copbio.2014.04.003.

Fernandes, P. 2010. "Enzymes in food processing: A condensed overview on strategies for better biocatalysts". *Enzyme Research* 2010: Article number 862537. https://doi.org/10.4061/2010/862537.

Gromiha, M. M., M. C. Pathak, K. Saraboji, E. A. Ortlund, and E. A. Gaucher. 2013. "Hydrophobic environment is a key factor for the stability of thermophilic proteins". *Proteins: Structure Function and Bioinformatics* 81, no. 4 (april): 715-21. https://doi.org/10.1002/prot.24232.

Haney, P. J., M. Stees, and J. Konisky. 1999. "Analysis of thermal stabilizing interactions in mesophilic and thermophilic adenylate kinases from the genus Methanococcus". *Journal of Biological Chemistry* 274, no. 40: 28453-58. https://doi.org/10.1074/jbc.274.40.28453.

Hao, L. L., P. C. Patel, A. H. Alhasan, D. A. Giljohann, and C. A. Mirkin. 2011. "Nucleic acid-gold nanoparticle conjugates as mimics of microRNA". *Small* 7, no. 22 (november): 3158-62. https://doi.org/10.1002/smll.201101018.

Horikoshi, K. 1998. "Barophiles: Deep-sea microorganisms adapted to an extreme environment". *Current Opinion in Microbiology* 1, no. 3 (june): 291-95. https://doi.org/10.1016/S1369-5274(98)80032-5.

Jafari, M., H. Abdollahi, S. Z. Shafaei, M. Gharabaghi, H. Jafari, A. Akcil, and S. Panda. 2019. "Acidophilic bioleaching: A review on the process and effect of organic-inorganic reagents and materials on its efficiency". *Mineral Processing and Extractive Metallurgy Review* 40, no. 2 (march): 87-107. https://doi.org/10.1080/08827508.2018.1481063.

Jatuwong, K., N. Suwannarach, J. Kumla, W. Penkhrue, P. Kakumyan, and S. Lumyong. 2020. "Bioprocess for production, characteristics, and biotechnological applications of fungal phytases". *Frontiers in Microbiology* 11 (february): Article number 188. https://doi.org/10.3389/fmicb.2020.00188.

Jayakumar, R., S. Jayashree, B. Annapurna, and S. Seshadri. 2012. "Characterization of thermostable serine alkaline protease from an alkaliphilic strain bacillus pumilus MCAS8 and its applications". *Applied Biochemistry and Biotechnology* 168, no. 7 (december): 1849-66. https://doi.org/10.1007/s12010-012-9902-6.

Johnson, D. B., and A. Schippers. 2017. "Recent advances in acidophile microbiology: Fundamentals and applications". *Frontiers in Microbiology* 8, no. MAR (march): 248. https://doi.org/10.3389/fmicb.2017.00428.

Johnson, E. A. 2013. "Biotechnology of non-Saccharomyces yeasts - The ascomycetes". *Applied Microbiology and Biotechnology* 97, no. 2 (january): 503-17. https://doi.org/10.1007/s00253-012-4497-y.

Jorda, J., and T. O. Yeates. 2011. "Widespread disulfide bonding in proteins from thermophilic archaea". *Archaea-an International Microbiological Journal* 2011: Article number 409156. https://doi.org/10.1155/2011/409156.

Kaur, G., and T. Satyanarayana. 2004. "Production of extracellular pectinolytic, cellulolytic and xylanoytic enzymes by thermophilic mould Sporotrichum thermophile Apinis in solid state fermentation". *Indian Journal of Biotechnology* 3, no. 4 (october): 552-57.

Korniłłowicz-Kowalska, T., and J. Bohacz. 2011. "Biodegradation of keratin waste: Theory and practical aspects". *Waste Management* 31, no. 8 (august): 1689-701. https://doi.org/10.1016/j.wasman.2011.03.024.

Ladenstein, R., and B. Ren. 2008. "Reconsideration of an early dogma, saying "there is no evidence for disulfide bonds in proteins from archaea"". *Extremophiles* 12, no. 1 (january): 29-38. https://doi.org/10.1007/s00792-007-0076-z.

Lewis, S. L. 2016. "The Paris Agreement has solved a troubling problem". *Nature* 532, no. 7599 (april): 283-83. https://doi.org/10.1038/532283a.

Li, S., X. Yang, S. Yang, M. Zhu, and X. Wang. 2012. "Technology prospecting on enzymes: Application, marketing and engineering". *Computational and Structural Biotechnology Journal* 2, no. 3 (september): e201209017. https://doi.org/10.5936/csbj.201209017.

Li, S. C., N. K. Goto, K. A. Williams, and C. M. Deber. 1996. "α-Helical, but not β-sheet, propensity of proline is determined by peptide environment". *Proceedings of the National Academy of Sciences of the United States of America* 93, no. 13 (june): 6676-81. https://doi.org/10.1073/pnas.93.13.6676.

Li, Z. Q., X. Pei, Z. Y. Zhang, Y. Wei, Y. Y. Song, L. N. Chen, S. A. Liu, and S. H. Zhang. 2018. "The unique GH5 cellulase member in the extreme halotolerant fungus Aspergillus glaucus CCHA is an endoglucanase with multiple tolerance to salt, alkali and heat: Prospects for straw degradation applications". *Extremophiles* 22, no. 4 (july): 675-85. https://doi.org/10.1007/s00792-018-1028-5.

Maruthamuthu, M., and J. D. van Elsas. 2017. "Molecular cloning, expression, and characterization of four novel thermo-alkaliphilic enzymes retrieved from a metagenomic library". *Biotechnology for Biofuels* 10, no. 1 (june): Article number 142. https://doi.org/10.1186/s13068-017-0808-y.

Merlino, A., I. Russo Krauss, I. Castellano, E. De Vendittis, B. Rossi, M. Conte, A. Vergara, and F. Sica. 2010. "Structure and flexibility in cold-adapted iron superoxide dismutases: The case of the enzyme isolated from Pseudoalteromonas haloplanktis". *Journal of Structural Biology* 172, no. 3 (december): 343-52. https://doi.org/10.1016/j.jsb.2010.08.008.

Mezzanotte, L., M. van 't Root, H. Karatas, E. A. Goun, and C. W. G. M. Lowik. 2017. "In vivo molecular bioluminescence imaging: New tools and applications". *Trends in Biotechnology* 35, no. 7 (july): 640-52. https://doi.org/10.1016/j.tibtech.2017.03.012.

Panja, A. S., B. Bandopadhyay, and S. Maiti. 2015. "Protein thermostability is owing to their preferences to non-polar smaller volume amino acids, variations in residual physico-chemical properties and more salt-bridges". *Plos One* 10, no. 7 (july): Article number e0131495. https://doi.org/10.1371/journal.pone.0131495.

Poppe, J. K., R. Fernandez-Lafuente, R. C. Rodrigues, and M. A. Z. Ayub. 2015. "Enzymatic reactors for biodiesel synthesis: Present status and future prospects". *Biotechnology Advances* 33, no. 5: 511-25. https://doi.org/10.1016/j.biotechadv.2015.01.011.

Rao, J. L. U. M., G. D. Y. Boorgula, and A. L. Leitão. 2011. "Fungal enzymes: Present scenario and future perspectives". In *Mycofactories*, 3-27. Bentham Science Publishers Ltd. https://doi.org/10.2174/978160805223311101010003.

Ribeiro, L. F., V. Amarelle, L. D. Alves, G. M. V. de Siqueira, G. L. Lovate, T. C. Borelli, and M. E. Guazzaroni. 2019. "Genetically engineered proteins to improve biomass conversion: New advances and challenges for tailoring biocatalysts". *Molecules* 24, no. 16 (august): Article number 2879. https://doi.org/10.3390/molecules24162879.

Rubingh, D. N. 1997. "Protein engineering from a bioindustrial point of view". *Current Opinion in Biotechnology* 8, no. 4 (august): 417-22. https://doi.org/10.1016/S0958-1669(97)80062-6.

Sansom, M. S. P. 1992. "Proline residues in transmembrane helices of channel and transport proteins: A molecular modeling study". *Protein Engineering, Design and Selection* 5, no. 1 (january): 53-60. https://doi.org/10.1093/protein/5.1.53.

Schmid, A., J. S. Dordick, B. Hauer, A. Kiener, M. Wubbolts, and B. Witholt. 2001. "Industrial biocatalysis today and tomorrow". *Nature* 409, no. 6817 (january): 258-68. https://doi.org/10.1038/35051736.

Schreck, S. D., and A. M. Grunden. 2014. "Biotechnological applications of halophilic lipases and thioesterases". *Applied Microbiology and Biotechnology* 98, no. 3 (february): 1011-21. https://doi.org/10.1007/s00253-013-5417-5.

Selvarajan, E., R. Veena, and N. Manoj Kumar. 2018. "Polyphenol oxidase, beyond enzyme browning". In *Microbial Bioprospecting for Sustainable Development*, 203-22. https://doi.org/10.1007/978-981-13-0053-0_10.

Sharma, A., Y. Kawarabayasi, and T. Satyanarayana. 2012. "Acidophilic bacteria and archaea: acid stable biocatalysts and their potential applications". *Extremophiles* 16, no. 1 (january): 1-19. https://doi.org/10.1007/s00792-011-0402-3.

Shrivastava, A., N. Shrivastava, and P. K. Singh. 2018. "Enzymes in pharmaceutical industry". In *Enzymes in food biotechnology: Production, applications, and future prospects*, 591-602. https://doi.org/10.1016/B978-0-12-813280-7.00034-7.

Singh, G., N. Capalash, K. Kaur, S. Puri, and P. Sharma. 2016. "Enzymes: Applications in pulp and paper industry". In *Agro-industrial wastes as feedstock for enzyme production: Apply and exploit the emerging and valuable use options of waste biomass*, 157-72. https://doi.org/10.1016/B978-0-12-802392-1.00007-1.

Srinivasan, M. C., and M. V. Rele. 1999. "Microbial xylanases for paper industry". *Current Science* 77, no. 1 (july): 137-42.

Stetter, K. O. 2011. "History of discovery of hyperthermophiles". In *Extremophiles Handbook*, 403-25. Springer Japan. https://doi.org/10.1007/978-4-431-53898-1_19.

Sunna, A., and P. L. Bergquist. 2003. "A gene encoding a novel extremely thermostable 1,4-beta-xylanase isolated directly from an environmental DNA sample". *Extremophiles* 7, no. 1: 63-70. https://doi.org/10.1007/s00792-002-0296-1.

Szilágyi, A., and P. Závodszky. 2000. "Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey". *Structure* 8, no. 5 (may): 493-504. https://doi.org/10.1016/S0969-2126(00)00133-7.

Takano, K., R. Higashi, J. Okada, A. Mukaiyama, T. Tadokoro, Y. Koga, and S. Kanaya. 2009. "Proline effect on the thermostability and slow unfolding of a hyperthermophilic protein". *Journal of Biochemistry* 145, no. 1 (january): 79-85. https://doi.org/10.1093/jb/mvn144.

Tompa, D. R., M. M. Gromiha, and K. Saraboji. 2016. "Contribution of main chain and side chain atoms and their locations to the stability of thermophilic proteins". *Journal of Molecular Graphics & Modelling* 64: 85-93. https://doi.org/10.1016/j.jmgm.2016.01.001.

Ueda, M., T. Goto, M. Nakazawa, K. Miyatake, M. Sakaguchi, and K. Inouye. 2010. "A novel cold-adapted cellulase complex from Eisenia foetida: Characterization of a multienzyme complex with carboxymethylcellulase, β-glucosidase, β-1,3 glucanase, and β-xylosidase". *Comparative Biochemistry and Physiology - B Biochemistry and Molecular Biology* 157, no. 1 (september): 26-32. https://doi.org/10.1016/j.cbpb.2010.04.014.

United Nations. 2008. *Kyoto Protocol Reference Manual*. United Nations Framework Convention on Climate Change.

———. 2015. *Adoption of the Paris Agreement*. Conference of the Parties on Its Twenty-First Session. https://digitallibrary.un.org/record/831039.

Vijayabaskar, M. S., and S. Vishveshwara. 2010. "Comparative analysis of thermophilic and mesophilic proteins using Protein Energy Networks". *Bmc Bioinformatics* 11, Suppll.1 (january): Article number S49. https://doi.org/10.1186/1471-2105-11-S1-S49.

Walsh, G. A., R. F. Power, and D. R. Headon. 1993. "Enzymes in the animal-feed industry". *Trends in Biotechnology* 11, no. 10 (october): 424-30. https://doi.org/10.1016/0167-7799(93)90006-U.

Wrigley, E. A. 2013. "Energy and the English Industrial Revolution". *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences* 371, no. 1986 (march): Article number 0568. https://doi.org/10.1098/rsta.2011.0568.

Wu, W. L., M. Y. Chen, I. F. Tu, Y. C. Lin, N. EswarKumar, M. Y. Chen, M. C. Ho, and S. H. Wu. 2017. "The discovery of novel heat-stable keratinases from Meiothermus taiwanensis WR-220 and other extremophiles". *Scientific Reports* 7, no. 1 (december): Article number 4658. https://doi.org/10.1038/s41598-017-04723-4.

Wu, X. R., Y. X. Wang, B. D. Tong, X. H. Chen, and J. H. Chen. 2018. "Purification and biochemical characterization of a thermostable and acid-stable alpha-amylase from Bacillus licheniformis B4-423". *International Journal of Biological Macromolecules* 109 (april): 329-37. https://doi.org/10.1016/j.ijbiomac.2017.12.004.

Yu, H. R., and H. Huang. 2014. "Engineering proteins for thermostability through rigidifying flexible sites". *Biotechnology Advances* 32, no. 2 (march): 308-15. https://doi.org/10.1016/j.biotechadv.2013.10.012.

Yu, H. R., Y. Zhao, C. Guo, Y. R. Gan, and H. Huang. 2015. "The role of proline substitutions within flexible regions on thermostability of luciferase". *Biochimica et Biophysica Acta-Proteins and Proteomics* 1854, no. 1 (january): 65-72. https://doi.org/10.1016/j.bbapap.2014.10.017.

Zaparty, M., and B. Siebers. 2011. "Physiology, metabolism, and enzymology of thermoacidophiles". In *Extremophiles Handbook*, 601-39. Springer Japan. https://doi.org/10.1007/978-4-431-53898-1_28.

Zhao, J., C. Guo, L. Zhang, and C. Tian. 2018. "Biochemical and functional characterization of a novel thermoacidophilic, heat and halo-ionic liquids tolerant endo-β-1,4-glucanase from saline-alkaline lake soil microbial metagenomic DNA". *International Journal of Biological Macromolecules* 118 (october): 1035-44. https://doi.org/10.1016/j.ijbiomac.2018.06.141.

Acknowledgments

This work was supported by the European Union [FEDER funds POCI/01/0145/FEDER/007728] and National Funds (Fundacão para a Ciência e Tecnologia and Ministério da Educação e Ciência − FCT/MEC) under the Partnership Agreement PT2020 [grant number UID/MULTI/04378/2019 and UIBD/50006/2020]; and the Fundação para a Ciência e a Tecnologia (FCT) [grant numbers PTDC/QUI-QFI/28714/2017, SFRH/BD/131968/2017].

4.2. QM/MM Study of the Reaction Mechanism of the Thermophilic Glucuronoyl Esterase for Biomass Treatment

Matilde F. Viegas [a], Rui P. P. Neves[a], Maria J. Ramos[a], and Pedro A. Fernandes[a]

[a] LAQV@REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

In this chapter, the catalytic mechanism of GE was described with atomistic detail, using QM/MM methodologies, and mutations, that might enhance the reaction rate of GE, were proposed.

Regarding the contributions to this work, Matilde F. Viegas performed the calculations with the guidance of Rui P. P. Neves and under the supervision of Pedro A. Fernandes, and co-supervision of Maria J. Ramos. Matilde F. Viegas wrote the first draft of the manuscript, and all authors have revised it through their contributions. This work was submitted to Catalysis, Science & Technology and it is currently under revision. The following content is an almost integral transcription of the submitted manuscript.

4.2.1. Abstract

A Hydrolysis of lignocellulosic biomass, composed of a lignin-carbohydrate-complex (LCC) matrix, is critical for producing bioethanol from glucose. However, current methods for LCC processing require costly and polluting industrial processes. Thermophilic enzymes can cope with the high-temperature conditions of industrial processes and are potential biocatalysts. The fungal *Thermothelomyces thermophila* glucuronoyl esterase (*Tt*GE) is a thermophilic enzyme and a promising biocatalyst that cleaves ester bonds in LCCs. Our study constitutes the first theoretical attempt to describe the glucuronoyl esterase catalytic mechanism, using QM/MM methods at the DLPNO-CCSD(T)/CBS:B3LYP/6-31G(d):AMBER level. In agreement with earlier experimental proposals, we confirmed that the catalytic cycle proceeds in two separate stages with ten stationary points. The rate-determining transition states are the nearly-degenerate TS1 and TS2, with a Gibbs barrier of *ca*. 17 kcal·mol^{-1}. Both have a zwitterionic nature that results from the proton interplay from the His346 base to either the Ser213-hydroxyl or the lignin leaving group and the rehybridization of the ester moiety of the substrate to an alkoxide. These transition structures are bridged by a highly endergonic tetrahedral intermediate. The hydrolysis is completed by a nucleophilic water molecule that breaks down the enzyme-hemicellulose adduct, priming the Ser213 for a new catalytic cycle. In both stages, the His346 of the canonical Ser-His-Asp triad triggers the nucleophilic attacks by Ser213 and the water molecule and resolves the resulting endergonic tetrahedral intermediates. The oxyanion hole, characteristic of esterases, was ensured by a conserved Arg214 through its backbone and sidechain. Our work further suggests that a mutation of Glu267 to a non-negative residue will decrease the energetic barrier *ca.* -5 kcal·mol^{-1}, improving the catalytic rate of *Tt*GE. Altogether, the results establish the groundwork for developing new strategies for improving LCC processing using *Tt*GE as a biocatalyst.

Keywords: LCC Processing, Lignocellulosic Biomass, Glucuronoyl Esterase, QM/MM, ONIOM, Enzymatic Catalysis.

4.2.2. Introduction

4.2.2.1. The biocatalytic potential of extremophiles for greener processes

The current state of energetic affairs demands a revitalization of energy production systems to sustain an ever-growing demand worldwide. The depletion of fossil fuel reserves, the political and socio-economic turmoil, and the indisputable effects of climate change have led to increased investment in renewable energy research [8]. . As climate change approaches an irreversible turning point, several countries have introduced targets to achieve net-zero emissions by 2050 [247]. Among several strategies, biofuels have become the hallmark in greener alternatives – biofuels are economically attractive, increasingly technologically feasible, and, above all, have reduced carbon emissions.

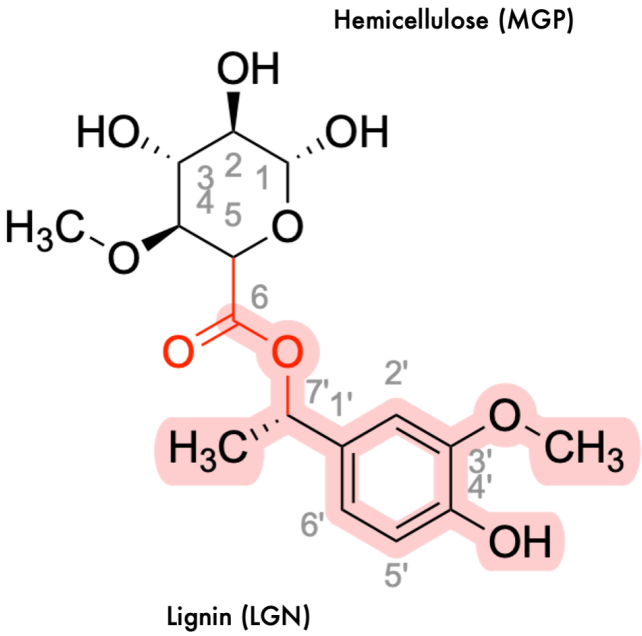
One route for biofuel production is bioethanol generation through glucose fermentation, the latter extracted from plant lignocellulosic biomass. Lignocellulosic biomass is composed of lignin (25 to 40%), and carbohydrates polymers - cellulose (30-50%), and hemicellulose (15-30%) [98–101]. To access fermentable sugar units in biomass, several linkages, such as esters, benzyl ethers, and phenyl glycosides, must be cleaved [1,102,105] to retrieve glucose units trapped within the complex matrix of (LCC)[102]. However, current methods for LCC material processing employ severe pretreatments, consuming both energy and polluting chemical reactants [103,104].

Such disadvantages call for cleaner alternatives, namely the consideration of carbohydrate enzymes, such as carbohydrate esterases, to efficiently catalyze LCC processing.

4.2.2.2. Glucuronoyl Esterase: an LCC degrading machine

The first carbohydrate esterase was discovered in 2006 by Spanikova [107] in a wood-rotting fungus called *Schizophyllum commune,* which presented cellulolytic activity. Enzymatic assays showed its ability for hydrolyzing unnatural, synthetic esters of 4-*O*-methyl-*D*-glucuronic acid and glucuronic acid, thus being named Glucuronoyl Esterase (GE) [107]. In 2007, the gene encoding GE was identified in *Trichoderma reesei*

(anamorph of *Hypocrea Jecorina*). A subsequent gene blast revealed the presence of Glucuronoyl Esterase coding genes in several microbial genomes. Its occurrence in different organisms led to the introduction of a new carbohydrate esterase (CE) family called CE15 in the Carbohydrate-Active Enzymes (CAZy) classification [108]. Members of the CE15 family are found in fungal and bacterial organisms, and all of them exhibit GE activity, cleaving esters of D-glucuronic acid. Interestingly, several bacterial CE15 enzymes were also found to be active on esters of galacturonoate, besides *D*-glucuronic acid [109], suggesting a broader substrate specificity. GEs belong to the α/β-hydrolase superfamily, which possesses a catalytic triad common with serine-type hydrolases: Glu/Asp-His-Ser [109–112]. Likewise, all CE15 enzymes exhibit the canonical catalytic triad except for a particular group of the CE15 family that exhibits a serine residue in place of the glutamic acid [113].



Scheme 3| MGPLGN substrate representative of LCC, representative of the most recalcitrant fraction of LCC – α ester – designed based on the MCU substrate of the *Tt*GE PDB ID: 4G4J X-ray structure [111]. MGPLGN is formed by a hemicellulose carbohydrate 4-O-methyl-β-D-glucopyranuronate (MGP) and a lignin (LGN) moiety (highlighted in red shade). The ester group is represented in red lines. This substrate is used as a representative minimal LCC in the study of the catalysis of *Tt*GE.

The first three-dimensional structure of a GE enzyme was obtained from the homologous over-expression of GE on the *T. reesei* in 2011 [110]. The structure displayed an α/β-hydrolase fold with an overall α/β/α-sandwich architecture. The twisted α-sheet is sandwiched between two layers of αβ-helices with the catalytic triad Ser-His-Glu exposed on the protein surface. A similar 3D structure was reported for GE from *Thermothelomyces thermophila* (which was previously named *Myceliophthora thermophila* and *Sporotrichum thermophile*; for clarity purposes, GE from *Thermothelomyces thermophila* will be addressed as *Tt*GE) [111]. In the same study, results demonstrated that a mutation of the catalytic serine to alanine in *Tt*GE abolished enzymatic activity, despite not affecting the integrity of its active site. This mutant was crystalized with an alternative substrate, methyl 4-*O*-methyl-β-*D*-glucopyranuronate (MCU) providing essential information about the active site organization when the substrate is bound [111]. Another interesting factor to consider is the positioning of the *Tt*GE's active site as surface-exposed, with its catalytic residues Ser213, Glu236 and His346 facing the solvent and exhibiting accessible surface areas of 27.1 Å^2, 2.7 Å^2 and 75.7 Å^2, respectively (Figure 10).

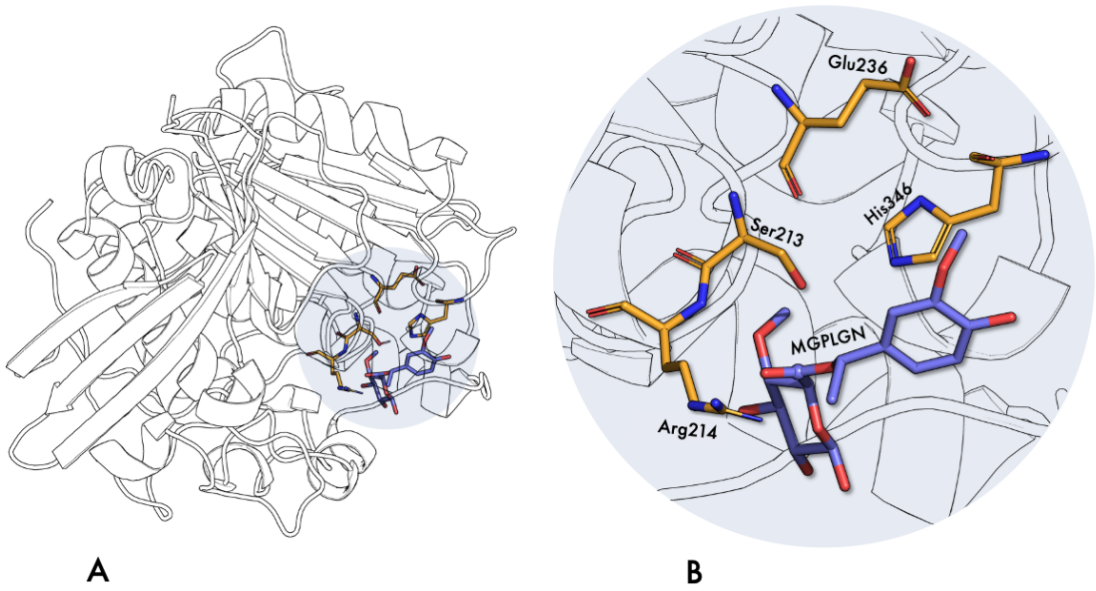
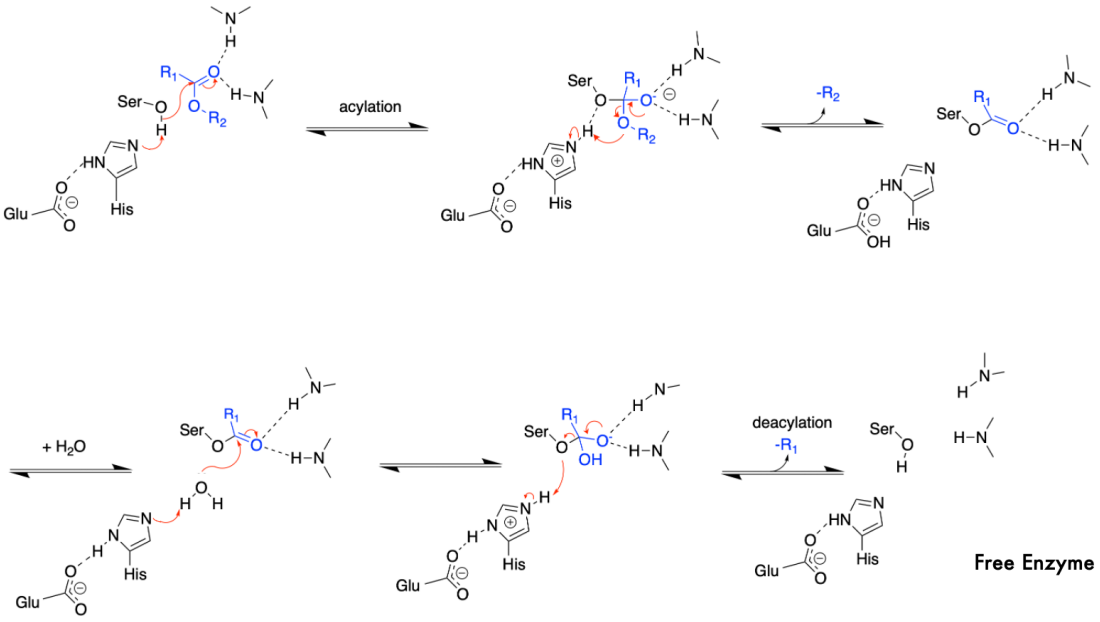


Figure 10| Representation of the *Tt*GE enzyme, obtained from the structure PDB ID: 4G4J. Catalytic residues are shown in stick representation (colored by element, yellow C atoms), and the docked substrate MGPLGN is shown in stick representation, colored by element with purple C atoms. The C6 carbon atom of the substrate MGPLGN is represented with a purple sphere.

This is consistent with the described role of the enzyme: cleaving bonds between large hemicellulose and lignin moieties. Finally, the catalytic triad is assembled in a rigid structure through a disulfide bond at the entrance of the active site, and a tight network of hydrogen bonds, forming a binding pocket that accommodates large lignocellulosic chains. It has also been found that GE enzymes integrate a modular system of enzymes, bi- or multi-modular, more often than not, in association with a carbohydrate-binding module CBM1, supporting its crucial role in plant cell-wall degrading mechanisms [111]. Other work supports the hypothesis that GE's expression appears to be inducible through different expression control systems [107,111,248].



Scheme 4| Schematic representation of the proposed canonical mechanism for esterase enzymes [249]. The initial substrate and its progression in the reaction mechanism is highlighted in blue.

A proposal for GE's enzymatic mechanism [249] follows that of a canonical esterase mechanism through a two- stage ping-pong bi-reaction involving an acyl-serine intermediate (Scheme 4). However, unlike the majority of GEs, where the catalytic triad is Ser-His-Asp [14], *Tt*GE contains a glutamic acid instead of an aspartic acid residue. Nevertheless, both residues appear to have the same mechanistic role – to stabilize the catalytic His346 residue through a hydrogen bond between its Nδ atom and Glu236 Oε1 [111].

The proposed canonical mechanism for esterase enzymes (Scheme 4) suggests that upon binding of the substrate at the active site, the nucleophilic attack of the catalytic Ser on the ester carbonyl carbon of the substrate yields a first tetrahedral intermediate. Deprotonation of the Ser by the catalytic His is key in this step. As the carbon-oxygen double bond forms, the R_2 portion of the substrate (equivalent to the lignin moiety, in the case of GE) is released into the solvent and protonated by the His, as the acyl-enzyme intermediate is formed [249]. Upon release of the R_2 leaving group, a hydrolysis reaction is initiated by the active site His, which acts again as a base to deprotonate the catalytic water molecule, prompting the nucleophilic attack of the water on the carbonyl carbon of the acyl-enzyme to form a second tetrahedral intermediate. Lastly, the alkoxide intermediate rehybridizes to carbonyl and releases the R1 portion (equivalent to hemicellulose), restoring the free enzyme state [249].

Kinetic data indicates that the presence of a 4-*O*-methyl group in the glucuronic acid increase the hydrolytic activity of the enzymes [107]. Additionally, it was shown that the glucuronoyl moiety has a more relevant role in recognizing the substrate than the lignin aromatic group [250]. In 2016, Bååth et al. established the first evidence of the enzymatic cleavage of lignocellulosic material by a GE, from *Acremonium alcalophilum* (AaGE1) [105]. Using size exclusion chromatography, their results displayed a reduced molecular mass of isolated lignin-carbohydrate complexes when incubated with recombinant GE, and they were able to identify an increase in carboxyl hydroxyl groups in solution. However, only recently, it was shown that GE has the unique ability to catalyze the removal of all glucuronoxylan associated with lignin through cleavage of ester linkages [251].

4.2.2.3. Nature's own catalysts as industrial workhorses

Stemming from the intrinsic characteristics of enzymes, industry and enzymes have become a long-time affair. Enzymes, nature's own catalysts, are fully biodegradable, highly selective, stereo and regiospecific, fast and efficient, and produce a low amount of by-products. The applicability of enzymatic catalyzers in industrial settings began in the 1960s, but the global market for industrial enzymes was already valued at USD

5.5 billion and is expected to reach USD 7.0 billion by 2023 [45]. To circumvent the need for polluting chemical catalyzers in harsh industrial settings, where mesophilic enzymes are rendered useless, extremophiles have garnered a reputation as a plausible alternative, seeing that these can withstand extreme conditions in terms of temperature, ionic strength, pressure, pH, toxicity and heavy metal content, captivating the interest of the scientific community [252].

Such interest can be confirmed in the *ca.* 2600 published papers between 2019 and 2021 (PubMed data). However, several technical challenges that need to be addressed to fully explore the potential of extremozymes, namely *i)* the isolation of high-quality DNA from environmental samples or *ii)* the effective detection of enzymatic activity in the conditions of the actual industrial process using hundreds of samples simultaneously. A second round of obstacles emerges in the fine-tuning of biocatalysts for industrial applications: firstly, to ensure that sensors and bioreactors can withstand the extreme conditions required for extremophile and extremozyme activity; and secondly, to ensure that by-product's accumulation will not hinder optimal extremophile growth.

Finally, and most importantly, current systems for extremophilic enzyme production depend on mesophilic hosts such as *Escherichia coli* or *Bacillus* spp., often causing a deficient production of enzymes due to improper folding and differences in the codon usage. As a result, to fully harness the potential of extremozymes, *i)* it is required to develop efficient molecular and culture tools to more effectively produce extremozymes, and *ii)* new genetic and protein engineering techniques must be developed to further improve the extremozymes action as biocatalysts in industrial settings [253].

With the current understanding of GE activity, this enzyme is now a potential thermophilic enzyme to be used in LCC processing. The QM/MM calculation of the present work resulted in the unveiled of GE's catalytic mechanism at the atomic level, disclosing its rate-determining steps. We have also identified residues surrounding the active site whose contribution could directly impact the GE's energetic barrier,

establishing the groundwork needed for mutagenesis studies to optimize GE's catalytic rate and efficiency for its application in industrial settings.

4.2.3. Computational Methods

4.2.3.1. Model Preparation

The computational model used in the present paper was based on the X-ray structure of GE from *T. thermophila, Tt*GE, containing a substrate analogue MCU, methyl 4-*O*-methyl-β-D-glucupyronuronate (PDB ID: 4G4J, 2.35Å resolution) [111]. We have chosen the X-ray structure PDB ID: 4G4J as the scaffold for our QM/MM model, as it presented the enzyme-ligand complex with the substrate analogue MCU located in its active site, in a quasi-reactive conformation, as opposed to other structures available at the time. The PDB ID: 4G4J structure presented a mutation in the catalytic residue Ser213Ala, corrected to the wild-type Ser213. Additionally, the His346 catalytic residue was modelled in the neutral form, while the Glu236 was modelled in its negatively charged form, agreeing with its solution pK_a. These protonation states are the only ones that allow for the classical chemical mechanism to be tested, as His346 should receive a proton from Ser213, and Glu236 should stabilize the formed imidazolium. In this work, we considered the protonation states for all amino acids at pH 7.0 according to PROPKA 3.1 [196,240].

The natural substrate of *Tt*GE is the lignocellulosic matter, an intricate matrix of lignin and hemicellulose encapsulating cellulose at its center. In the employed X-ray structure [111], a substrate analogue named MCU, methyl 4-*O*-methyl-β-*D*-glucopyranuronate, was present at the active site. The MCU molecule served as a template to model our substrate, as we replaced the methyl moiety of the substrate MCU (representative of the hemicellulose portion of the natural substrate) with a representative monomer of lignin 4-(1-hydromethyl)-2-methylphenol (LGN), resulting in our final ligand MGPLGN (Scheme 3). The target of GE is the ester bond found between the carbohydrate MeGluPy and LGN moieties of the MGPLGN ligand. Modeling of the substrate molecule was performed with GaussView5 [242]. Docking was

performed with the molecular docking software GOLD (Genetic Optimization Ligand Docking) [254] and the MCU molecule present in the considered X-ray structure [111] served as the template for the docking procedure. Modeling of the substrate analogue formed by a carbohydrate 4-*O*-methyl-β-*D*-glucopyranuronate (MeGluPy) and a lignin (LGN) moiety was followed by a conventional molecular dynamics simulation (cMD) to ensure the quality of the resulting structure. As previously mentioned, the PDB ID:4G4J crystal structure employed in modelling our structure displayed a mutation of Ser213 to Ala213. Charavgi *et al.* mentioned that the absence of the hydroxyl group did not affect the architecture of the active site and that the hydrogen bonds formed between the Ala213 backbone atoms (N and O) and nearby residues, as well as the contacts with His346 and Glu236, were maintained. The architecture of the active site is further sustained by the disulfide bond formed between Cys347 and Cys212 [111]. Thus, employing a crystal structure in a quasi-reactive conformation and proceeding with physical constraints applied to the active site residues (Cys212, Ser213, Arg214, Glu236, His346 and Cys347) and to the substrate, the modeling protocol adopted in the study ensures the quality of the resulting structure.

The parameters for the substrate MGPLGN were determined from an optimized structure at the HF/6-31G(d) level of theory. These were obtained with the Antechamber tool, as implemented in AMBER12 [179], using the GAFF force field [243] to derive the intramolecular and Lennard-Jones parameters, and the RESP (Restrained Electrostatic Potential) [244] method to derive atomic charges at the HF/6-31G(d) level of theory. The enzyme was parameterized using the AMBER FF99SB force field [199,245], while ions and hydrogen atoms were added using the XLeap module of Amber12 software [179]. TIP3P water molecules [181] were added within 16.0 Å from the modelled system, to fill a rectangular box.

cMD calculations were performed using the AMBER 12 software [179], after an energy minimization of water molecules, followed by hydrogen covalent bonds, then side-chains, and finally of the entire system. The four minimization steps were applied for model optimization and to avoid tensions and clashes derived from previous modeling.

To constrain bonds involving hydrogen atoms in all cMD calculations, the SHAKE algorithm [246] was employed, and an integration step of 2 fs was used. Annealing was performed from 0 to 310 K during 40 ps at constant volume, followed by another 40 ps at 310 K in the *NVT* ensemble. During this phase, the protein backbone remained constrained. The above-mentioned constraints employed harmonic potential force constants of 50 kcal·mol^{-1} Å^{-2}. cMD productions were then performed to analyze the conformations of the active site region in an *NPT* ensemble at 1 bar, and 310 K. The Berendsen barostat and the Langevin thermostat were used throughout the simulations.

Throughout the QM/MM study of the GE catalytic mechanism, we performed three separate protocols for the cMD production phase. Before QM/MM modelling of the reactant state, a cMD of the system comprising the reactant state (56037 atoms) – named cMD1 for clarity purposes – was run. The cMD1 production phase protocol followed was: i) 6 ns with protein backbone and substrate constrained, and unconstrained water molecules; ii) 6 ns with protein backbone constrained, except for the substrate and active site residues Cys212, Ser213, Glu236, Arg214, His346 and Cys347; iii) followed by 24 ns of unrestrained cMD.

The leaving group (protonated LGN) is expected to leave the active site at the end of the acylation step, and the catalytic water molecule is required at the active site. To further sample the conformational space, we performed cMD2 with the system resulting from the acylation step (56,037 atoms), in which the LGN group establishes a hydrogen bond (2.15 Å) with the Nε2 of the His346. We carried out four different 42 ns cMD production phase simulations, using different starting velocities, with i) a 2 ns period of constrained acyl-enzyme backbone; ii) 40 ns, with no constraints applied. The LGN group left the active site in all simulations, diffusing into the solvent. Hence, for the QM/MM modelling of the deacylation step, we chose to model the acylated enzyme assuming that the LGN leaving group quickly diffuses into the bulk solvent (56,013 atoms). The cMD3 production phase was carried out for 82 ns: i) an initial 2 ns period where the protein and substrate were kept constrained, allowing the water molecules to relax; ii) 40 ns period of constrained protein backbone except active site

residues of the acyl-enzyme (Cys212, Ser213 bonded to the MGP carbohydrate moiety, Glu236, Arg214, His346 and Cys347), iii) followed by 40 ns, with no constraints applied. During this cMD simulation, we also observed the existence of water molecules buried at the active site by visual inspection, using the criterium that the water molecule must be present within a 3 Å radius from both the C6 of the acyl-enzyme and the Nδ1 of the catalytic His346.

4.2.3.2. QM/MM Model

The QM/MM enzyme:substrate model for the acylation step of the reaction (Stage 1) included the whole enzyme and the substrate, as well as all the waters within *i)* a 6 Å distance from the active site as defined by the residues Cys212, Ser213, Glu236, Arg214, His346 and Cys347 and *ii)* a 3 Å distance from the protein surface, in a total of 8,178 atoms, with an overall charge of -6 and a singlet spin multiplicity. Initial coordinates for the model were assigned after 12 ns of cMD production (cMD1), in which the enzyme backbone was kept fixed, except for the active site. This structure was chosen because it is very similar to the X-ray enzyme structure, which by virtue of representing an average over all proteins in the crystal, is the best conformational option to perform adiabatic mapping QM/MM calculations [145]. The QM layer comprised 132 atoms, including the active site residues side chains (Ser 213, His 346 and Glu236), the substrate, the two cysteine residues forming the disulfide bond nearest to the active site (Cys212, Cys347) and the Arg214, necessary for the oxyanion conformation (as seen in Figure 13). Additionally, four water molecules were included in the DFT layer (Figure 11, Stage 1). The system was further divided into a frozen and a free region: all atoms beyond 20 Å radius from the catalytic site were kept frozen, alongside all water molecules except those included in the DFT layer [255].

For the deacylation step of the reaction (Stage 2), the QM/MM model was built as for the Stage 1 except that it no longer included the LGN moiety of the substrate, resulting in an 8,052 atoms system. The QM layer comprised 110 atoms, including the substrate (now covalently bonded to Ser213 and without LGN), Ser213, His346, Glu236, Cys212

and Cys347, Arg214 and three water molecules (Figure 11, Stage 2). Again, the same conditions were applied to freeze the atoms in the system.

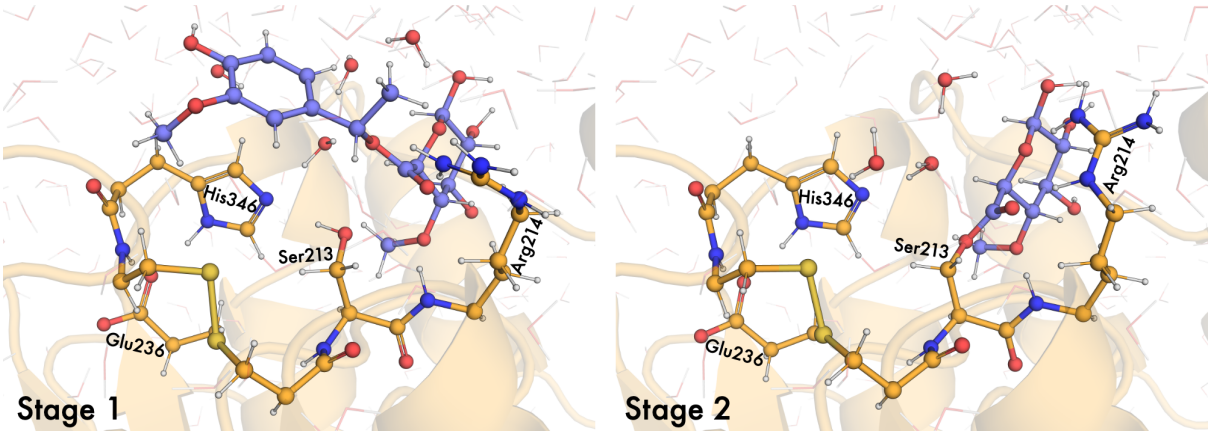


Figure 11| Overall QM/MM models for Stage 1 and Stage 2 of the reaction mechanism. Transparent cartoon representation of the GE protein, as well as the transparent stick representation of water molecules, depict the considered MM layer. The atoms represented in ball and stick belong to the QM layer.

4.2.3.3. QM/MM calculations and reaction coordinates

The QM/MM calculations were performed using the electrostatic embedding scheme of the ONIOM method [256], available in Gaussian 09 [257]. The link-atom approach [179] was used to truncate bonds at the QM boundary by adding hydrogen atoms to fill the valence of capped residues in the QM layer.

Initial linear transit scans and subsequent full, free geometry optimizations of all stationary states were performed at the B3LYP/6-31G(d):FF99SB level of theory. A vibrational analysis confirmed the absence of imaginary frequencies in every reactant, intermediate and product state for each minimized structure, corroborating their status as the real minima. Regarding the fully optimized transition states, we confirmed the presence of a single imaginary vibrational frequency corresponding to a maximum only along with the reaction coordinate of the PES. The reaction path was also studied for each transition state by following the reaction's intrinsic reaction coordinate (IRC) [258] of the reaction.

Single-point calculations were conducted to refine the QM/MM energies at the extrapolated DLPNO-CCSD(T)/CBS level. The latter calculations were applied to a

smaller QM layer: 55 atoms in Stage 1 of the reaction, and 49 atoms in Stage 2, both upon filling valences with hydrogen atoms. This approach has shown superior accuracy for activation energies of enzymatic reactions at an affordable computing time in a recent benchmarking study [259,260]. Hence, Hence, the final energy calculations for all stationary points were performed at the DLPNO-CCSD(T)/CBS:B3LYP/6-31G(d):FF99SB//B3LYP/6-31G(d):FF99SB level. Thermal corrections to the Gibbs energy at the standard ambient temperature and pressure were calculated at the B3LYP/6-31G(d):FF99SB level using the particle-in-the-box/rigid rotor/harmonic oscillator approximations for the translational, rotational and vibrational components of energy and entropy.

4.2.4. Results and discussion

The sections include a detailed discussion of the results of the computational calculations regarding the QM/MM study of the reaction mechanism of GE and geometry changes in the stationary states of the reaction. For clarity purposes, it will be subdivided into three distinct subsections: first, a study of the prevalence of water molecules in the active site, which is crucial for the catalytic cycle, followed by a description of the results for the first and the second stages of the reaction. The following section describes a mutagenesis study to identify active site neighboring residues that contribute to the limiting step energetic barrier. The final subsection will dive into the detailed comparison of the *Tt*GE and its fellow CE15 family members.

4.2.4.1. Water: the key for the catalysis

As a serine-type hydrolase, the proposed mechanism for GE involves the canonical catalytic triad and a catalytic water required for the second part of the reaction [249]. From the close examination of the electron density of the X-ray crystal structure PDB ID: 4G4J [111], with a 2.35 Å resolution, we observe the presence of two water molecules, which are further confirmed by the corresponding electron density map, indicating that these water molecules are likely to be conserved in the immediacy of the active site residues and the ligand (3.76 Å and 4.16 Å from the MCU ligand). The

presence of water molecules at the active site is expected as: *i)* the active site is exposed to the solvent, and it possesses several polar residues (active site Ser213 and His346, plus Lys217 and Glu267) that establish hydrogen bonds with surrounding waters; *ii)* being a serine-type hydrolase, a water molecule is required to resolve the deacylation step of the reaction, where the catalytic Ser213 is regenerated, and the product is released.

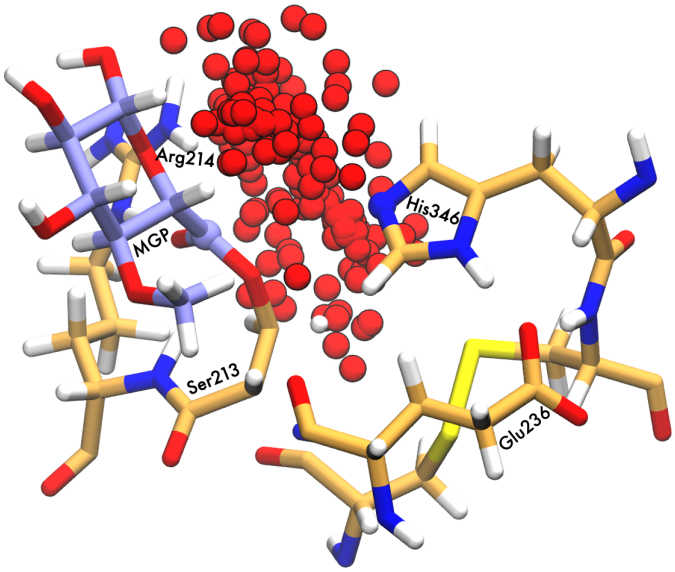


Figure 12| Water molecules that can be found in the immediacy of the active site (within 5 Å of both carbon C6 of the glycosyl-enzyme and nitrogen Nε of His346), every 1 ns over the course of an unrestrained 40 ns cMD simulation. Water molecules are represented in red spheres, surrounding the active site residues, in yellow colored stick representation, and the glycosyl MGP, in purple colored stick representation. Carbon C6 of the glycosyl-enzyme and nitrogen Nε of the His346 are highlighted in sphere representation.

Visual inspection of the cMD shows that water molecules readily access and occupy the free space within the active site, favouring the hypothesis that the reaction catalysed by GE requires water molecules for catalysis (Figure 12). A close inspection of the interactions established at the active site shows a tight network of hydrogen bonds between water molecules and neighbouring residues. The binding pose of the substrate at the active site was conserved for the whole cMD, namely through

conserved interactions between the uronic acid portion and Arg214, Gln259, Glu267, Trp310, as well as with solvent water molecules (as seen in Figure 16A).

4.2.4.2. First Reaction Stage: Formation of the acyl-serine enzyme intermediate

Once the substrate is located at the solvent-exposed active site, the activity of the GE enzyme is dependent on its catalytic triad Ser213-His346-Glu236, where the Ser213 provides the essential nucleophile for the hydrolysis to occur.

In the geometry-optimised reactant state (REACT, Figure 13), the Ser213 Oγ is aptly oriented towards the ester carbon (C6 of MGPLGN) for the nucleophilic attack: at a distance of 2.78 Å, and with a 92.49° angle between the Ser213 Oγ, the C6 and the carbonyl oxygen, in line with the Bürgi-Dunitz angle referred for other serine-proteases [261], and extended to esterases [262]. The Glu236 is hydrogen-bonded to the Nδ of the His346 (1.67Å), with the latter's Nε hydrogen-bonded to a water molecule (1.89Å) and to the Ser213 hydroxyl (2.06 Å). The Arg214 is hydrogen-bonded to the carbonyl oxygen of the ester, through both its backbone amine (2.53 Å) and guanidine (2.07 Å) groups, forming an oxyanion hole proposed to stabilize the growing negative charge that develops at the carbonyl oxygen during the nucleophilic attack.

We observed that the nucleophilic attack by Ser213 occurs in a concerted, yet asynchronous fashion: as Ser213 moves towards carbon C6, His346 deprotonates the Ser213 Oγ to trigger the nucleophilic addition to carbon C6, and a transient covalent tetrahedral intermediate is formed - Intermediate 1 (INT1, Figure 13).

The transition state (TS1, Figure 13) is observed when the Ser213 Oγ is at 2.03 Å distance from carbon C6, and the His346 is positively charged. The hydrogen bonds between Arg214 and the ester carbonyl oxygen, which is now partially anionic, become tighter, stabilizing TS1 through a classical oxyanion hole effect. As the reaction progresses from REACT to TS1, the distance between the Arg214 backbone amine and carbonyl oxygen shorten from 2.53 Å to 2.15 Å, whilst the Arg214 guanidine amine and carbonyl oxygen distance increases from 2.07 Å to 2.26 Å. Additionally, we observed that the ionic interaction between the Nδ of His346 and Oε1 of Glu236, both part of the catalytic triad, shortens from 1.67 Å to 1.60 Å, stabilizing the now positively-charged His346.

INT1 formed when Ser213 Oγ forms a stable covalent bond with the substrate. The distance between the Ser213 Oγ and carbon C6 shortened from 2.04 Å at TS1 to 1.51 Å at INT1. The hydrogen bond between the Nε of His346 and the nearby water molecule, present in the REACT, are lost at TS1 and INT1 (from 1.89 Å at REACT to 2.72 Å at TS1, to 3.04 Å at INT1). Instead, the water molecule establishes a new hydrogen bond with the oxo oxygen (O7) of the LGN moiety and promotes a rearrangement of the neighbouring hydrogen bond network with three adjacent water molecules. TS1 exhibited a Gibbs activation energy ($\Delta G_1^{\ddagger}$) of 15.7 kcal·mol^{-1} and led to an intermediate product (INT1) with a Gibbs reaction energy of 11.6 kcal·mol^{-1}. Vibrational frequency analysis revealed an imaginary frequency dominated by the bond stretching between the nucleophilic Oγ of Ser213 and the carbon C6 of the substrate (123.1i cm^{-1}), characteristic of nucleophilic addition, confirming the nature of the transition state. In particular, we were only able to characterize the TS1 in the explicit presence of water molecules near the leaving group, which suggests that solvent may play a role in stabilizing the leaving LGN.

The final step of Stage 1 is the deprotonation of neighbouring His346 by the LGN moiety of the substrate, followed by the formation of the acyl-enzyme covalent adduct (Figure 13). As the His346 Nε is deprotonated by the leaving oxo oxygen of the LGN moiety, a transition state (Transition State 2, TS2, Figure 14) is observed. The distance between carbon C6 and the oxo oxygen increases from 1.52 Å at INT1 to 1.93 Å at TS2, while the His346 remains protonated, with the Nε proton at 1.71 Å from the leaving oxo oxygen of the LGN moiety. TS2 exhibited a Gibbs activation energy ($\Delta G_2^{\ddagger}$) of 16.8 kcal·mol^{-1} and led to an intermediate product (INT2) with a Gibbs reaction energy of 3.1 kcal·mol^{-1}, both in relation to REACT. The vibrational frequency analysis revealed an imaginary frequency corresponding to the reaction coordinate, confirming the nature of the transition state TS2 (99.36i cm^{-1}). The analysis of the imaginary frequency highlights the cleavage of the C-O bond with the LGN leaving group and the protonation by His346; both reactions are accompanied by a rearrangement of the H-bond network by the nearby waters from the leaving LGN to theHis346 (Figure 13).

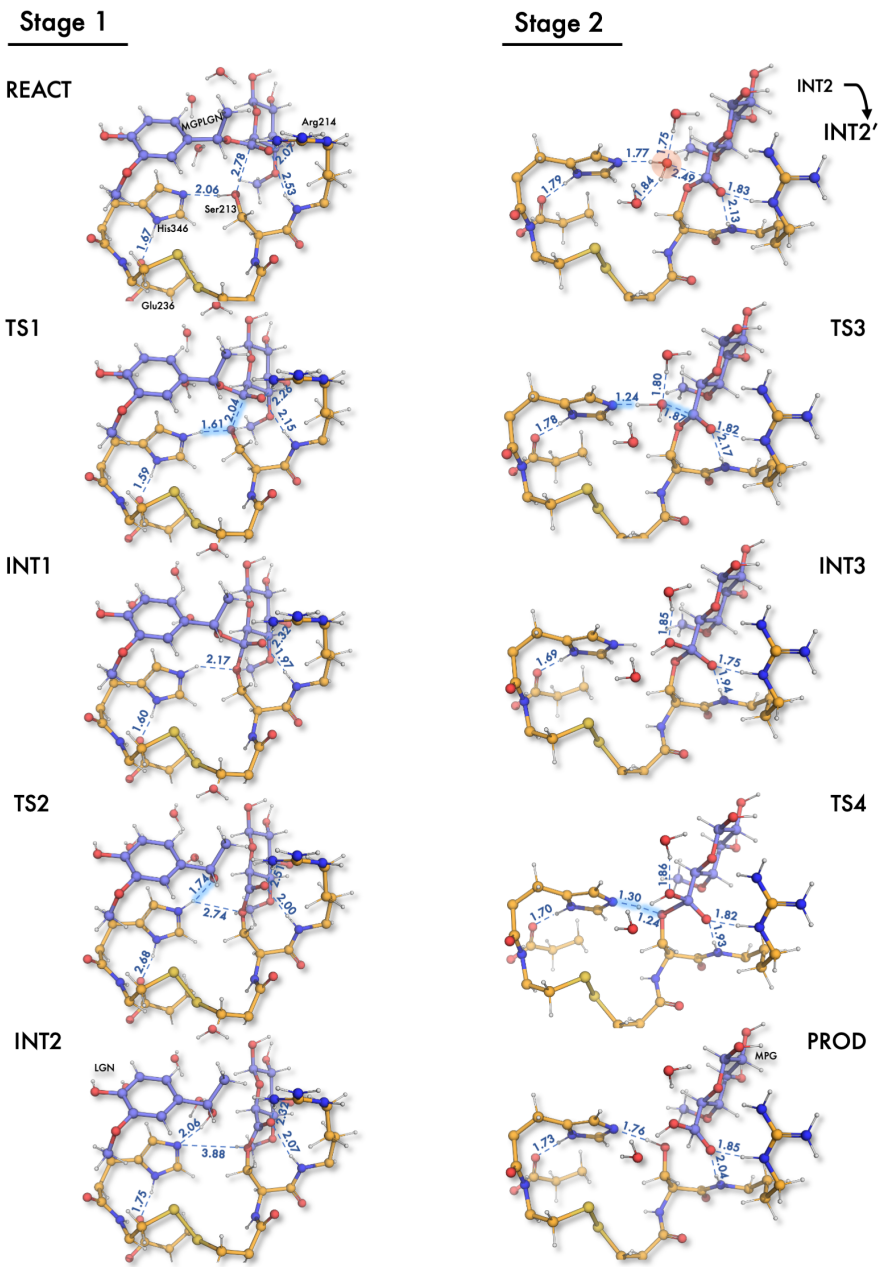


Figure 13| Stationary points of the *Tt*GE reaction. QM layer represented in ball and stick, and the substrate colored in purple. In the INT2' stage, the nucleophilic bulk water molecule is highlighted with a red-colored circle. In dashes, the most relevant Interatomic distances that were followed along the QM/MM calculations, obtained at the B3LYP/6-31G(d):FF99SB level.

The final INT2 corresponds to the protonated LGN leaving group via His346 deprotonation, which means that the dissociation of the tetrahedral intermediate and the late deprotonation of His346 by the leaving LGN are concerted asynchronous

steps. After the completion of this step, cMD simulations have shown that the LGN portion of the substrate quickly dissociates from the active site (within 2 ns), which is subsequently filled by water molecules (similarly to what is depicted in Figure 3). It is from this state that the Stage 2 of the catalytic cycle begins.

4.2.4.3. Second Reaction Stage: Deacylation and Serine Regeneration

The current proposal for Stage 2 consists of the deacylation and regeneration of the enzyme - a close reverse of Stage 1 of the reaction. An activated water molecule should act as a nucleophile towards carbon C6 of the enzyme-MeGluPy adduct [249]. As water is a crucial element of the reaction, we investigated the availability of water molecules within the active site by cMD simulations (Figure 12). The residence rate of the water molecules at the area between the His346 and C6 of the acyl-enzyme was *ca.* 24%, which translates to a binding energy of water in this region of *ca.* +1 kcal·mol^{-1}.

Stage 2 starts from the INT2' state (Figure 13), where the oxygen of the nucleophilic water molecule is oriented towards carbon C6 at a 2.49 Å distance, exhibiting a 112.55° angle with carbon C6, nucleophilic water molecule oxygen, and nitrogen Nε of His346. As the active site is located at the interface between the enzyme and solvent, the nucleophilic water forms hydrogen bonds with two other water molecules (1.75 Å and 1.84 Å). The Arg214 is hydrogen-bonded to the C6-carbonyl oxygen through its guanidine and backbone amine groups (1.83 Å and 2.13 Å, respectively) and to oxygen O6 in the hemicellulose moiety (1.87 Å), anchoring it and contributing to the correct positioning of the ester for catalysis. We also observe a hydrogen bond between Nδ of His346 and Oε1 of Glu236, with a 1.79 Å distance.

As the nucleophilic water approaches carbon C6 of the acyl-enzyme, the His346 deprotonates the water molecule, resulting in a highly nucleophilic hydroxide-like species primed to covalently bind C6, as observed in the transition state 3 (TS3, Figure 13). In TS3, the water oxygen is at 1.87 Å from the C6, while the water proton lies in between the Nε of His346 and the water oxygen (1.24 Å and at 1.29 Å). Hence, it is clear that the reaction proceeds through a concerted and synchronous transition state.

Once the hydroxide ion forms a covalent bond with carbon C6, the tetrahedral intermediate (INT3, Figure 13) is formed. Just as observed in Stage 1, as the tetrahedral intermediate is formed, the negative charge building in the carbonyl oxygen is stabilized by an oxyanion hole with the neighboring Arg214: 1.95 Å (backbone amine) and 1.75 Å (guanidine).

Vibrational frequency analysis revealed an imaginary frequency corresponding to the reaction coordinate confirming the nature of the transition state TS3 (944.43i cm^{-1}). The imaginary vibration associated with TS3 is characterized by a pronounced stretching of the HO bond of the nucleophilic water molecule that is accompanied by the rehybridization of carbon C6 upon the nucleophilic attack. A Gibbs activation energy ($\Delta G_3^{\ddagger}$) of 12.1 $kcal \cdot mol^{-1}$ was determined and led to an intermediate state (INT3) with a Gibbs reaction energy of 5.2 $kcal \cdot mol^{-1}$ in relation to the INT2'.

In subsequent reaction step, the bond between oxygen Oγ of the Ser213 and carbon C6 breaks down in the path between INT3 and transition state 4 (TS4, Figure 13). The resulting negatively charged Ser213 Oγ readily deprotonates the positively charged His346. In TS4, the His346 proton is found at 1.24 Å from the Ser213 Oγ, and at 1.30 Å from the Nε of His346. Consequently, the tetrahedral intermediate disassembled, and the Ser213-His346 dyad is regenerated, while the resulting carbohydrate product (PROD, Figure 13) is released, getting the enzyme GE ready for a new catalytic cycle. The transition state TS4 exhibited a Gibbs activation energy ($\Delta G_4^{\ddagger}$) of 13.0 $kcal \cdot mol^{-1}$ relatively to the INT2' and led to the PROD with a Gibbs reaction energy of 10.2 $kcal \cdot mol^{-1}$. The vibrational frequency analysis revealed an imaginary frequency corresponding to the deprotonation of His346 by Ser213, thus confirming the nature of the transition state TS4 (1046.29i cm^{-1}).

4.2.4.4. Energetic profile for the GE catalytic mechanism

A burst kinetics study conducted by Weiner *et al.* presented data substantiating the hypothesis that the rate-limiting step for esterases occurs before forming the acyl-enzyme intermediate [263]. In more recent years, the reaction rates reported for GE are

approximately 115.9 min^{-1} with the substate trans-3-phenyl-2-propen-1-yl D-glucopyranosyluronate and 166.4 min^{-1} with 3-phenyl-1-propyl D-glucopyranosyluronate as substrate [248]. The Gibbs energy barrier estimated using Transition State theory is thus *ca.* 18 $kcal \cdot mol^{-1}$ for the hydrolysis of the ester bond between 4-O-methyl-D-glucuronic acid residues of glucuronoxylans and aromatic alcohols of lignin.

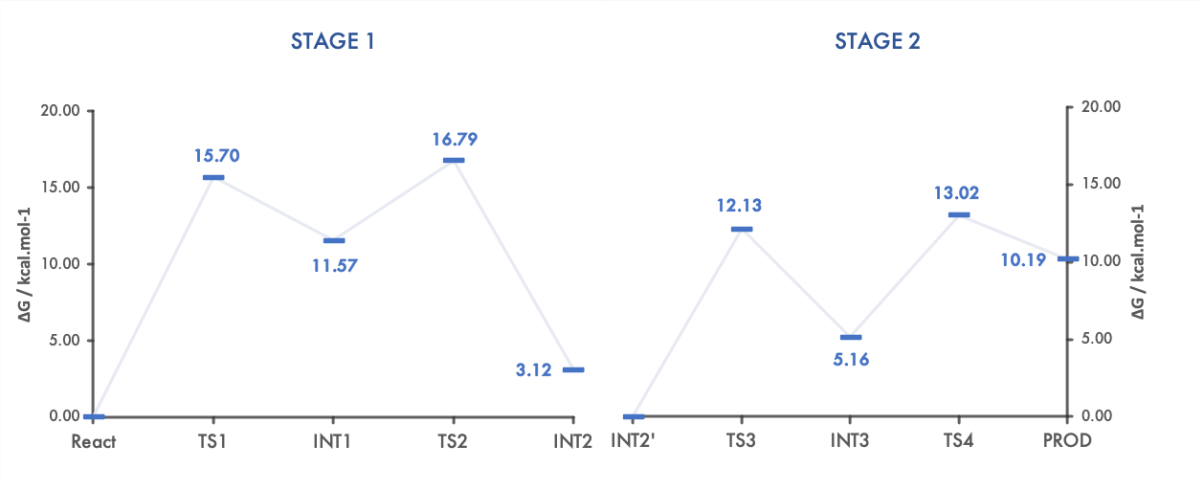
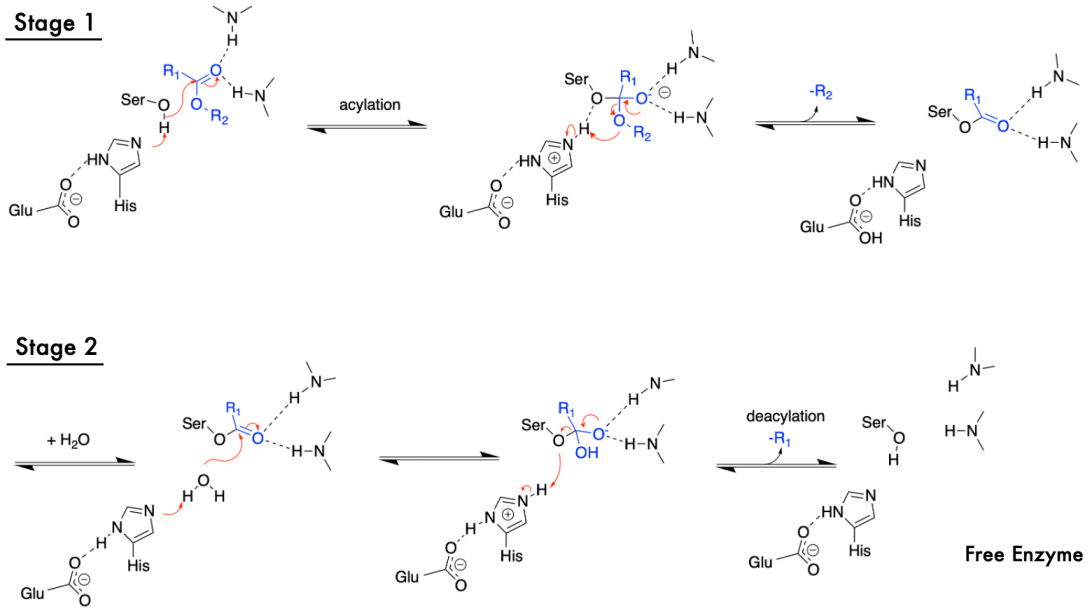


Figure 14| Gibbs Energy profile of the global GE-catalyzed carbohydrate hydrolysis, at the DLPNO-CCDS(T)/CBS:FF99SB//B3LYP/631G(d):FF99SB level of theory. The interruption between the INT2 and INT2' states, of both Stages 1 and 2, was generated by the removal of the free LGN leaving group.

In the present work, the step before the formation of the acyl-enzyme intermediate (REACT->TS1) has a Gibbs energy barrier of 15.7 $kcal \cdot mol^{-1}$ and a Gibbs reaction energy of 11.6 $kcal \cdot mol^{-1}$, whilst the deprotonation of the His346 by the LGN moiety of the substrate to form the acyl-enzyme covalent adduct (INT1->TS2) has a Gibbs energy barrier of 5.2 $kcal \cdot mol^{-1}$, thus corresponding to a Gibbs energy barrier of 16.8 $kcal \cdot mol^{-1}$ measured from REACT (Figure 14), in line with the experiments. Since connecting the Gibbs energy profiles of the two stages would require knowing the Gibbs energy associated with the dissociation of the LGN portion of the substrate and solvation of the active site, which would imply complex calculations that, additionally, do not provide very accurate results, the Gibbs energy profiles of each stage of the cycle are analyzed separately. We cannot absolutely deny that there is not an even higher barrier at Stage 2 (measured in relation to REACT) if INT2' is considerably

higher in Gibbs energy than INT2. However, the cMD simulations indicate that both the exit of the LGN and the solvation of the active site carry a negligible kinetic cost because both phenomena were frequently observed on an ns-timescale, whereas the catalysis by GE takes place beyond the ms-timescale. Hence, our results strongly suggest that the formation of the enzyme-substrate adduct is the rate-limiting stage of the catalytic cycle of GE. TS1 and TS2 are too close in energy to identify if one of them, and which one of them, is rate-limiting, so we assume that, within the accuracy of the calculations, both can be rate-limiting. The resulting mechanism is depicted in Scheme 5.



Scheme 5| Schematic representation of Glucuronoyl Esterase reaction mechanism. The initial substrate and its progression in the reaction mechanism is highlighted in blue.

4.2.4.5. Transition-State Macromolecular Dipole Stabilization

The rate-limiting transition states are the nearly-degenerate TS1 and TS2. Both have a zwitterionic nature, with a positive charge at His346 and a negative charge delocalized between the serine hydroxyl and the substrate ester carbonyl (TS1), and between the leaving group oxygen and the substrate ester carbonyl (TS2).

It has been previously suggested that charged residues surrounding the active site have a significant impact on the energetic barriers of the enzymatic reaction, as their interactions can potentially stabilize or destabilize the charge rearrangements that happen in the course of the rate-limiting reaction step [226]. Specifically, negative residues closer to His346 than to the ester carbonyl and hydroxyl of Ser213 should stabilize TS1 more than the reactants and thus lower the barrier. However, the opposite should be the truth for positive residues. Concerning TS2, negative residues closer to His346 than to the ester carbonyl and leaving group oxygen should stabilize it more than the reactants, lowering the barrier, with the opposite being true for positive residues.

To explore this hypothesis, we calculated the energetic contribution for the REACT, TS1, and TS2 states of every single residue within 10 Å of the active-site region. In addition, we evaluated the energetic contributions of mutated residues for both reaction steps of Stage 1 because they have very similar energy barriers. The results are shown in Figure 15.

Figure 15A indicates that *Tt*GE is remarkably optimized - almost no residue raises the barrier. All but one (Glu267) are contributing to lowering the activation barrier. Interestingly, *Tt*GE is poor in ionic residues, with only 6 charged residues evenly distributed between Nε His346 and the ester carbonyl oxygen of the substrate MGPLGN. Similar results were observed for both reaction steps. From the negatively charged residues (colored in red), the Glu267 is closer to the ester carbonyl oxygen of the substrate MGPLGN, and its contribution to the barrier is +6 kcal·mol^{-1}. Noteworthy is Lys217, also closer to the ester carbonyl oxygen of the substrate MGPLGN contributed to the barrier by -10 kcal·mol^{-1}. Glu305 is closer to Nε of His346 than to the ester carbonyl oxygen, and thus it contributed to a decrease in the barrier of -5 kcal·mol^{-1}. These results are in agreement with previous work done by our group [203,226]. These contributions can be significant, in particular if their effect is cooperative. The structural positioning of the aforementioned residues (Lys217, Glu305 and Glu267) is highlighted in Figure 15B.

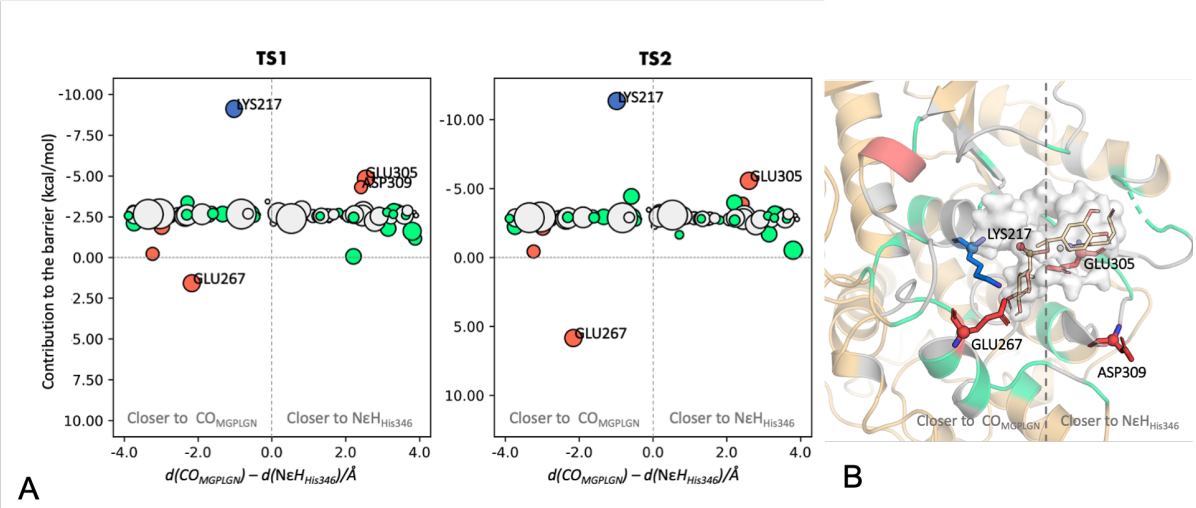


Figure 15| Per-residue activation energy contribution as a function of the difference in the distances of the centre of mass of the sidechain (or the charged group for the charged residues) of the residue to the ester carbonyl oxygen of the substrate MGPLGN and Nε of His346, in Å. Larger circles represent bulkier amino acids. Positively charged, negatively charged, polar, and hydrophobic residues are colored, respectively, in blue, red, green, and gray. The energy contributions were evaluated at the ONIOM(B3LYP/6-311+G(2d,2p):AMBER) level. B) Representation of the residues within 10 Å of the active site of TtGE. Positively charged, negatively charged, polar, and hydrophobic residues are colored, respectively, in blue, red, green, and grey. Labelled residues lead to the most significant changes in the activation energy. The dashed grey line separates residues closer to ester carbonyl oxygen of MGPLGN than to NεH of His346.

The obvious path to optimize the enzyme kinetics would be to mutate Glu267 by a similar, non-negative residue, such as glutamine, or to introduce positive/negative residues in the regions where they benefit the enzyme kinetics, the first closer to His348 than to Glu305 and the second in the opposite way. Nevertheless, the introduction of charged residues in the enzyme scaffold is challenging, as it may lead to conformational changes that affect their predicted position, and thus further *in vitro* testing will be necessary to confirm the effect of the mutations.

4.2.4.6. Comparison with members of the CE15 family - active site rigidity and substrate binding

With the detailed picture of the catalytic reaction mechanism by GE, in particular the nature of the rate-limiting step and the active site molecular determinants contributing for GE catalysis, these results can be further extended to other members of the CE15

family, as they comprise some of the most promising enzymes regarded for lignocellulosic material degradation.

As such, in this section, we analyze the structural similarities and differences between GE enzymes from *Opitutus terrae* (anaerobic bacteria, grows optimally between 10-37°C [264]), *Cerrena unicolor* (white rot fungus, grows optimally at 28°C [265]), and *Thermothelomyces thermophila* (white rot fungus, grows optimally at 45-50°C [266]), highlighting structural scaffolding such as disulfide bonds, hydrogen bond networks, and conserved residues at the active site, in light of our mechanistic results.

Our results indicated that critical residues of the rate-limiting step of GE (the formation of enzyme-substrate adduct by the nucleophilic Ser213) are the well-known His346, which activates the Ser213 for the nucleophilic attack, and the Arg214, which alone forms the oxyanion hole arrangement required to stabilize the resulting tetrahedral intermediate. In addition, the availability of water was also shown to provide the polar environment required to stabilize the tetrahedral intermediate. Upon substrate binding, the Arg214 residue shows conformational changes, corroborating its role in forming the oxyanion hole we characterized in the QM/MM study. It has been suggested that the arginine residue, despite its variation in position in the bacterial and fungal structures known to date, has a transition state stabilizing role. The conservation in >95% (26 of 27) of characterized CE15 sequences [108] supports its crucial role in catalysis. In the continuous electron density of the crystal structure of the catalytically active fungal *Cerrena unicolor* GE (*Cu*GE) covalent complex with the aldotetrauronic acid (PDB ID: 6RV8) [267], it is possible to observe that the Arg residue establishes hydrogen bonds with the covalent-bound substrate, indicating the formation of the oxyanion hole. Alignment of the aforementioned *Cu*GE covalent complex structure with the TS1 conformation of *Tt*GE shows a perfect alignment of active site residues Ser213-His346-Glu236 and the glucuronoyl moieties of the substrates in both structures. Finally, the backbone of the Arg residue responsible for the oxyanion hole formation is aligned in both structures, substantiating the role of Arg in both the substrate recognition and the formation of the oxyanion hole [267].

Throughout the cMD simulations of the GE:MGPLGN complex, the predominant interactions established between the MGPLGN substrate and neighboring residues

anchored the substrate in the active site for 36 ns. The prevalent interactions observed were: 1) two hydrogen bonds between the MGPLGN ester carbonyl oxygen and the hydrogens of the backbone amide and guanidine group of Arg214 (2.88±0.84 Å; 4.41±0.90 Å), and 2) two hydrogen bonds established between one hydrogen of the MGP moiety (from the OH on the C1 of the sugar) and Gln316 side chain oxygen atoms (3.98±0.47 Å; 2.55±1.01 Å). Further analysis of the interactions shows that the uronic acid part of the MGPLGN substrate interacts with residues Trp310, Gln259 and Lys217, all of which were previously described in the co-crystallization of an inactive *St*GE2 mutant with methyl 4-*O*-methyl-D-glucopyranuronate [111].

Kinetic data consistently suggests that the presence of a 4-*O*-methyl group at the glucuronic acid increases hydrolytic activity by the GE [40,107,250]. Furthermore, crystal structures of *Cu*GE with aldouronic acids of three and four sugar units – mimicking its natural LCC substrates - have deepened the knowledge around substrate recognition in GE enzymes [267]. By analyzing the MCU substrate present in the crystal structure of *Tt*GE (PDB ID: 4G4J) and the MGPLGN ligand used in the present study, we observe that the glucuronoyl moiety is responsible for establishing the hydrogen bonds that anchor the substrate in the active site (Figure 16). Further analysis of three GE structures co-crystalized with ligands (Figure 17) indicate that the glucuronic acid moiety is key in the H-bond positioned deep into the active site pocket, despite differences in the electrostatic surface potential between *Tt*GE, *Cu*GE and bacterial *Ot*GE. This is in agreement with previous studies, which indicated that the glucuronoyl moiety is more important for substrate recognition than the aromatic group [107,250].

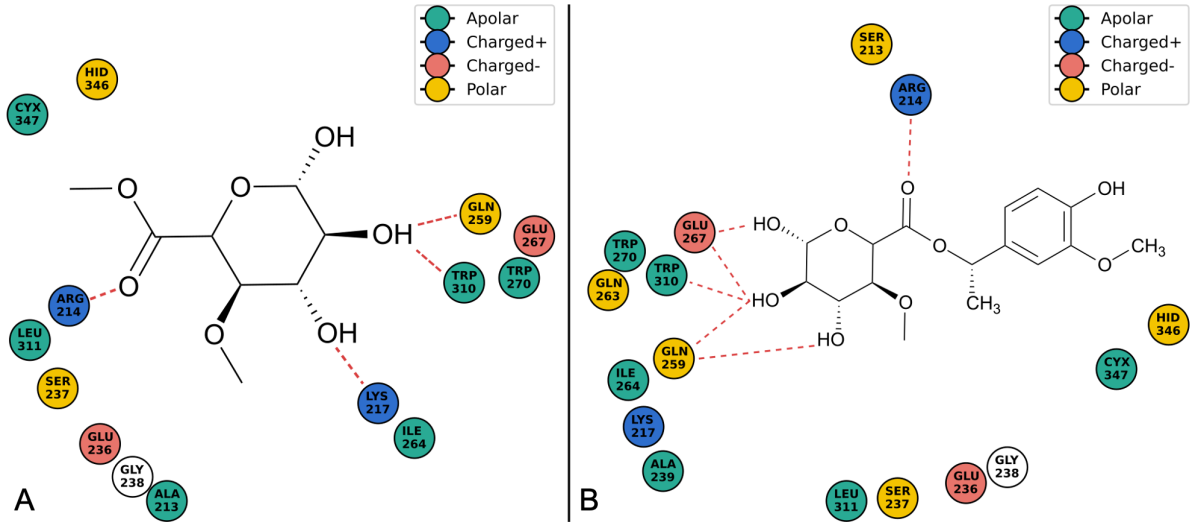


Figure 16| 2D Diagrams of the H-Bond interactions between the ligand and neighboring residues present at the Active Site, displaying the predominance of interactions established between the uronic acid moiety and the protein. A) the Crystal Structure of TtGE (PDB ID:4G4J) and its co-crystalized MCU ligand; B) At the Active Site of the TtGE with the MGPLGN ligand used for the study of the Catalytic Reaction at Stage 1.

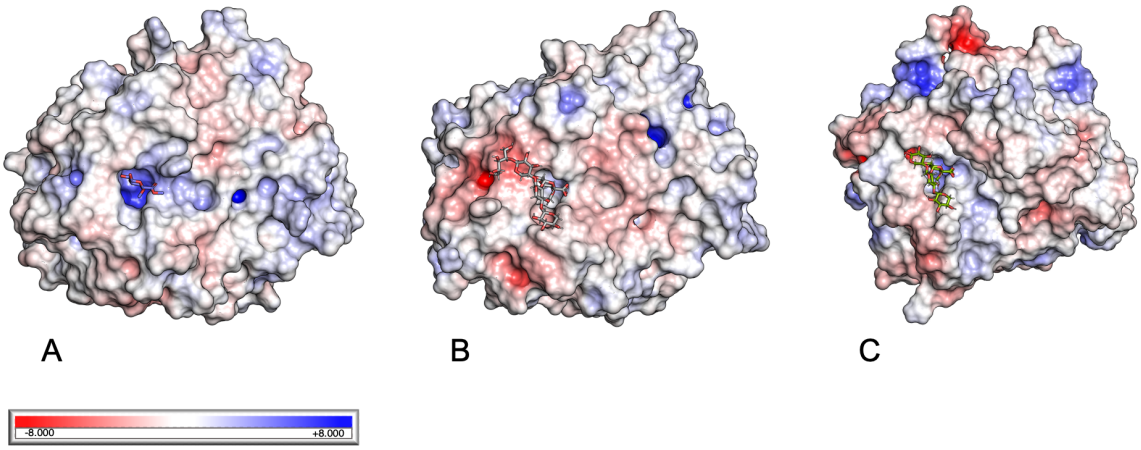


Figure 17| Electrostatic Surface Potential representation of fungal and bacterial GEs. A. Fungal TtGE (PDB ID: 4G4J) with docked substrate; B. Fungal CuGE with docked substrate (PDB ID:6RV9, chain B). C) Bacterial OtGE with docked substrate (PDB ID: 6T0I).

As members of the α/β-hydrolase superfamily, CE15 enzymes display extremely robust active sites. In the case of GEs from different domains of life, the conservation seen across is remarkable, showing little variation between structures determined at different conditions (pH, active *vs.* inactive, apo *vs* ligand-bound).

The active site of CE15 enzymes is accessible at the protein's surface, substantiating its role as an LCC degrading enzyme with large substrates. In the case of fungal CE15 (such as *Tt*GE and *Cu*GE), the presence of the disulfide linkage anchoring the catalytic residues His and Ser suggests that the presence of the two cysteine residues is key for the higher rigidity in its solvent-exposed active site. The importance of the three disulfide linkages found in *Tt*GE (and conserved in the fungal GE enzymes known to date) is a question worth addressing in future research further to enhance the resistance of thermophilic enzymes in industrial contexts. Another important feature contributing to the rigidity of the active site is the short hydrogen bond at the entrance of the active site between Gln235 and Glu267 (1.8 Å) in *Tt*GE and residues Glu320 and Glu324 (2.5 Å) in *Cu*GE (PDB ID:6RV8) [267].

Interestingly, disulfide bonds and the aforementioned hydrogen bonds are absent in bacterial GE structures known to date. As bacterial GEs are known for the broader substrate specificity relative to their fungal homologs, the lack of rigidity conferring features support this hypothesis. Recent studies in the bacterial GE MZ0003 described the existence of a novel sub-domain, which forms a deeper positively charged substrate pocket [112]. Moreover, in place of the Gln235 found at the *Tt*GE, there is an Arg residue in both bacterial GE MZ0003 and *Ot*GE and, as observed in *Ot*GE, this Arg residue interacts with the substrate, anchoring next to the active site. Altogether, these factors are likely responsible for and indicate that diversity of activity may exist among unexplored clades of CE15 enzymes.

4.2.5. Conclusions

In this study, we explored the reaction mechanism of glucuronoyl esterases, which have garnered attention in the past decade as clear candidates for eliminating lignin from hardwood by cleavage of the ester bond between 4-O-methyl-*D*-glucuronic acid residues of glucuronoxylans and aromatic alcohols of lignin [102,268]. We have outlined

the Gibbs energy profile for both proposed steps of glucuronoxylan degradation by GE and the molecular determinants of the enzyme catalysis, using computational methods and QM/MM models.

We confirmed that the catalytic cycle proceeds in two separate stages, confirming the previous hypothesis in the literature, with a total of ten stationary points. The rate-limiting transition states are the nearly-degenerate TS1 and TS2. Both have a zwitterionic nature, with a positive charge at His346 and a negative charge delocalized between the serine hydroxyl and the substrate ester carbonyl (TS1) to hold a highly endergonic tetrahedral intermediate that is quickly resolved to produce the enzyme-substrate adduct and between the leaving group oxygen and the substrate ester carbonyl (TS2). Such a result is in excellent agreement with previous experimental studies. As *Tt*GE is an esterase with the active site exposed to the solvent, its deacylation in stage 2 of the reaction was carried out by a nucleophilic water molecule from the surrounding solvent. A nucleophilic bulk water molecule is deprotonated by the His346 and covalently binds the C6 atom of the Ser213-glycosyl, resulting in a second endergonic tetrahedral intermediate. As the bond between oxygen Oγ of the Ser213 and carbon C6 breaks down, the resulting negatively charged Ser213 Oγ readily deprotonates the positively charged His346, getting the active site of GE ready for a new catalytic cycle. As expected with esterases, the oxyanion hole formation was observed in both catalytic stages and was guaranteed by an Arg214 conserved in about 95% of the CE15 enzyme family.

We have also mapped a path to optimize the enzyme kinetics of *Tt*GE, suggesting the mutation of Glu267 by a similar, non-negative residue such as glutamine, plus further mutations in the enzyme periphery to create a macrodipole that stabilizes the zwitterionic transition state.

The results are finally discussed in other enzymes of the CE15 family, identifying important patterns of enzyme reactivity and function for subsequent enzyme engineering studies. As the CE15 enzymes have been noted to vary in their substrate specificity, in particular the bacterial type, there is a tremendous biotechnological potential to be explored in this enzyme family. The conclusions drawn in the present study detail the intricacies of the catalysis carried out by GE enzymes, laying the groundwork for further enhancement of these nature's catalysts. The presented

enzymes are promising candidates as auxiliary enzymes to improve saccharification of plant biomass.

4.3. Outcomes and Future Prospects

LCC complexes are responsible for the resistance of plant structures, but they also hinder the successful utilization of lignocellulosic biomass. Therefore, the interest in glucuronoyl esterases has intensified in the past decade due to their potential in degrading LCC complexes. *Tt*GE is a glucuronoyl esterase enzyme found in a thermophilic fungus capable of degrading lignocellulosic biomass. With our work, we were able to determine the catalytic mechanism of this enzyme with an atomistic detail, and determine the energetic barrier and limiting step. Furthermore, our *in silico* mutagenesis assays have revealed mutations that can potentially decrease the energetic expenditure, increasing its enzymatic efficiency.

The published literature review, introduced in subchapter 4.1., delves into the subject of thermophilic enzymes and the structural features that convey their capacity to operate in temperatures above 40 °C, providing a broader context to the understanding of *Tt*GE and its relevance. Throughout our research, a preliminary assay on the relevance of the disulfide bonds for TtGE's active site integrity and catalysis suggested that disulfide bonds play a role in maintaining a catalytic-prone conformation of the active site. This is in agreement with previous studies, which indicate that a decrease in flexibility is essential to the evolutionary drive toward a selective and proficient catalysis [269]. To further explore this question, a comparison between the thermophilic GE and its mesophilic counterpart could be made by assessing the importance of their disulfide bonds. This could be achieved by performing extensive cMD simulations of thermophilic GE with its cysteine residues either forming disulfide bonds or in their neutral states (mutant), alongside a similar cMD of its mesophilic counterpart. This would provide data to evaluate the stability of the enzyme structure, through backbone RMSd, and of its active site, and through RMSd of residues in the active site and those within close distance, 3.5 Å. The key distances of the reaction mechanism of GE would also be assessed and compared to the distances established in our study of *Tt*GE's catalytic mechanism. Estimation of activation and reaction energies could also be made through QM/MM simulations of

the rate-limiting step of GE mechanism with each of the *Tt*GE's models (mesophilic, thermophilic, and thermophilic mutant), similarly to what was done on the Transition State Macromolecular Dipole Stabilization on subchapter 4.2.4.5. This evaluation of disulfide bonds' contribution to the energetic barrier could also contribute to developing more catalytically efficient enzymes for industrial processes.

As a thermophilic, lignocellulose-processing enzyme, its utilization in an industrial environment could contribute to replacing polluting chemicals currently employed in lignocellulosic biomass processing. Combining *Tt*GE with other enzymes, such as α-glucuronidase, is possible to process LCC complexes: α-glucuronidase would initiate processing by the initial release of longer xylan-oligosaccharides from the LCC linkage followed by the release of 4-O-methyl-glucuronic acid by the GE to fully yield carbohydrates from lignin [113].

CHAPTER 5. Snake venom as a reservoir for therapeutic agents

A natural reservoir for bioactive molecules is venom, which is defined as *a poisonous substance secreted by animals such as snakes, spiders, and scorpions and typically injected into prey or aggressors by biting or stinging* by the Oxford Dictionary. Natural evolution has led to over 220,00 species producing some type of venom [270], with snakes and scorpions as the most well-known venomous animals for the general public.

Snake venom contains a mixture of 20-100 bioactive toxins, with the vast majority being peptides and proteins [271], many of which work synergically when injected into the prey. The composition of snake venom varies extensively between species and even between animals of the species, with sex, age, and type of prey influencing the ratio of compounds in the venom [272–275].

In the scope of the study of snake venom and its enzymes, we also collaborated with a team of experimental scientists based in Ecuador and Brazil. We provided the resolved structure of a basic secreted phospholipase A_2, $sPLA_2$ (D49), from the venom of the snake *Agkistrodon piscivorus leucostoma* (ApTX-II), with the docked synthetic substrate NOBA (4-nitro-3-octanoyloxybenzoic acid). This structure was modeled using homology modeling and subjected to MD simulations with explicit solvent. NOBA and active-site water molecules spontaneously adopted stable positions and established interactions in full agreement with the reaction mechanism proposed for the physiological substrate, suggesting that NOBA hydrolysis is an excellent model to study phospholipid hydrolysis. The enzymatic activity of the ApTX-II was assessed through *in vivo* and *in vitro* assays. Results show that its enzymatic activity depends on the presence of Ca^{2+} in the medium. Comparing the primary structure with other $sPLA_2$s X-ray structures revealed that ApTX-II presented all the structural reasons expected for a basic $sPLA_2$s.

5.1. Structural, Enzymatic and Pharmacological profile of ApITX-II - A Basic sPLA$_2$ (D49) isolated from the *Agkistrodon piscivorus leucostoma* snake venom

Letícia M. Resende[a], José R. Almeida[b], Tatiana A. Guaraca-Medina[b], Cleopatra A. da S. Caldeira[c,d], Matilde F. Viegas[e,f], Andreimar M. Soares[c,d], Maria J. Ramos[e,f], Pedro A. Fernandes[e,f], Sergio Marangoni[a], Saulo L. Da Silva[e,g]

[a] Department of Biochemistry and Tissue Biology, Institute of Biology, Campinas State University (UNICAMP), Campinas, SP, Brazil

[b] IKIAM - Universidad Regional Amazonica, Km 7 Via Muyuna, Tena, Napo, Ecuador

[c] Biotechnology Laboratory for Proteins and Bioactive Compounds from the Western Amazon, Oswaldo Cruz Foundation Rondonia, Porto Velho, Rondonia, Brazil

[d] Oswaldo Cruz Foundation (FIOCRUZ), CEBio, Fiocruz Rondonia and Federal University of Rondonia, Porto Velho, RO, Brazil.

[e] LAQV/Requimte, University of Porto, Rua do Campo Alegre s/n, Porto, Portugal

[f] Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Rua do Campo Alegre s/n, Porto, Portugal

[g] College of Biochemistry and Pharmacy, Faculty of Chemical Sciences, University of Cuenca, Cuenca, Azuay, Ecuador

---

Resende, L. M., Almeida, J. R., Guaraca-Medina, T. A., **Viegas, M. F.**, Soares, A. M., Ramos, M. J., Fernandes, P. A., Marangoni, S., & Da Silva, S. L. (2021). ***Structural, enzymatic and pharmacological profiles of ApITX-II - A basic sPLA$_2$ (D49) isolated from the* Agkistrodon piscivorus leucostoma *snake venom.*** In International Journal of Biological Macromolecules, 175, 572-585. https://doi.org/10.1016/j.ijbiomac.2021.01.187

5.2.1. Abstract

A basic $sPLA_2$ (D49) from the venom of the snake *Agkistrodon piscivorus leucostoma* (AplTX-II) was isolated, purified and characterized. We determined the enzymatic and pharmacological profiles of this toxin. AplTX-II was isolated with a high level of purity through reverse phase chromatography and molecular exclusion. The enzyme showed pI 9.48 and molecular weights of 14,003 Da. The enzymatic activity of the AplTX-II depends on Ca^{2+}, pH and temperature. The comparison of the primary structure with other $sPLA_2$s revealed that AplTX-II presented all the structural reasons expected for a basic $sPLA_2$s. Additionally, we have resolved its structure with the docked synthetic substrate NOBA (4-nitro-3-octanoyloxybenzoic acid) by homology modeling, and performed MD simulations with explicit solvent. Structural similarities were found between the enzyme's modeled structure and other snake $sPLA_2$ X-Ray structures available in the PDB database. NOBA and active-site water molecules spontaneously adopted stable positions and established interactions in full agreement with the reaction mechanism proposed for the physiological substrate, suggesting that NOBA hydrolysis is an excellent model to study phospholipid hydrolysis.

Keywords: Basic $sPLA_2$ D49; Agkistrodon piscivorus leuscostoma; Molecular Modeling; NOBA (4-nitro-3-octanoyloxybenzoic acid)

International Journal of Biological Macromolecules 175 (2021) 572–585

Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

journal homepage: http://www.elsevier.com/locate/ijbiomac

Structural, enzymatic and pharmacological profiles of AplTX-II - A basic sPLA$_2$ (D49) isolated from the *Agkistrodon piscivorus leucostoma* snake venom

Letícia M. Resende [a], José R. Almeida [b], Tatiana A. Guaraca-Medina [b], Matilde F. Viegas [e,f], Andreimar M. Soares [c,d], Maria J. Ramos [e,f], Pedro A. Fernandes [e,f], Sergio Marangoni [a], Saulo L. Da Silva [e,f,g,*]

[a] Department of Biochemistry and Tissue Biology, Institute of Biology, Campinas State University (UNICAMP), Campinas, SP, Brazil
[b] IKIAM - Universidad Regional Amazonica, Km 7 Via Muyuna, Tena, Napo, Ecuador
[c] Biotechnology Laboratory for Proteins and Bioactive Compounds from the Western Amazon, Oswaldo Cruz Foundation Rondonia, Porto Velho, Rondonia, Brazil
[d] Oswaldo Cruz Foundation, FIOCRUZ, Rondonia and Federal University of Rondonia, UNIR, Porto Velho, RO, Brazil
[e] LAQV/Requimte, University of Porto, Rua do Campo Alegre s/n, Porto, Portugal
[f] Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Rua do Campo Alegre s/n, Porto, Portugal
[g] Faculty of Chemical Sciences, University of Cuenca, Cuenca, Azuay, Ecuador

ARTICLE INFO

Article history:
Received 27 July 2020
Received in revised form 25 January 2021
Accepted 28 January 2021
Available online 30 January 2021

Keywords:
Basic sPLA$_2$ D49
Agkistrodon piscivorus leuscostoma
Molecular modeling
NOBA (4-nitro-3-octanoyloxy benzoic acid)

ABSTRACT

A basic sPLA$_2$ (D49) from the venom of snake *Agkistrodon piscivorus leucostoma* (AplTX-II) was isolated, purified and characterized. We determined the enzymatic and pharmacological profiles of this toxin. AplTX-II was isolated with a high level of purity through reverse phase chromatography and molecular exclusion. The enzyme showed pI 9.48 and molecular weight of 14,003 Da. The enzymatic activity of the AplTX-II depended on Ca^{2+} pH and temperature. The comparison of the primary structure with other sPLA$_2$s revealed that AplTX-II presented all the structural reasons expected for a basic sPLA$_2$s. Additionally, we have resolved its structure with the docked synthetic substrate NOBA (4-nitro-3-octanoyloxy benzoic acid) by homology modeling, and performed MD simulations with explicit solvent. Structural similarities were found between the enzyme's modeled structure and other snake sPLA$_2$ X-Ray structures, available in the PDB database. NOBA and active-site water molecules spontaneously adopted stable positions and established interactions in full agreement with the reaction mechanism, proposed for the physiological substrate, suggesting that NOBA hydrolysis is an excellent model to study phospholipid hydrolysis.

© 2021 Elsevier B.V. All rights reserved

1. Introduction

Snake venoms are complex secretions produced in a specialized gland that can be injected in prey or in defensive circumstances (such as when humans are bitten) to produce deleterious effects including immobilization, paralysis, death, and digestion [1,2]. Snakebite envenoming in humans occurs all over the world, mainly in rural areas of Africa, Asian and Latin America, and the World Health Organization considers it a priority neglected tropical disease since June 2017 [3,4]. Accidents with snakes, principally with species from the *Elapidae* and *Viperidae* families, are responsible for 421,000 to 1,841,000 envenomings, leading to 94,000 deaths and 300,000 severe (often chronic) physical consequences [5–8].

Snake venoms possess a high number of distinct kinds of molecules (most of which proteins), such as phospholipases A_2, metalloproteinases, L-amino acid oxidases, sarafotoxins, disintegrins, among others [9–20]. Most of these substances are the subject of research regarding the relationships between their structures and their pharmacological activities [21,22]. The principal characteristic of toxins present in venom secretions is that they act specifically on their pharmacological targets. This characteristic has helped researchers and pharmaceutical companies to produce many essential and efficient drugs used nowadays, most of them used for blood disease therapy [23–27,29]. However, the literature has shown that there are different toxins present in snake venoms that have interesting antitumor, antihypertensive, antimicrobial and antiparasitic activities, as well as other important activities that have the potential to help produce new drugs [23–27,30–38].

Phospholipases A_2 (PLA$_2$ - EC 3.1.1.4) are lipolytic enzymes that hydrolyze the ester bond at the sn-2 position of glycerophospholipids, producing arachidonate acid and lysophospholipids [34]. Secreted phospholipases A_2 (sPLA$_2$) are probably the most studied protein

* Correponding author at: Faculty of Chemical Sciences, University of Cuenca, Cuenca, Azuay, Ecuador.
E-mail address: saulo.da.silva@ucuenca.edu.ec (S.L. Da Silva).

https://doi.org/10.1016/j.ijbiomac.2021.01.187
0141-8130/© 2021 Elsevier B.V. All rights reserved

components of snake venoms and display a variety of pharmacological activities, such as myotoxic, pre- and post-synaptic neurotoxic, anticoagulant, edematogenic, hemorrhagic, hypotensive, hemolytic and bactericidal [9,15,31,34,39–44].

Several important works have helped to understand the generic aspects of the catalytic mechanisms of $sPLA_2$s, showing that phospholipid hydrolysis directly involves the key active site residues H48 and D49. Additionally, residues D99, Y52, and Y73, a water molecule, and a Ca^{2+} ion (which determines that $sPLA_2$s are calcium-dependent) also participate in the complete reaction mechanisms. Nevertheless, a detailed description of these enzymes' reaction mechanisms, with an atomic level step-by-step characterization of intermediates and transition states, is still missing [31,45–51].

Agkistrodon piscivorus leucostoma (*A. p. leucostoma*, Western cottonmouth rattlesnake) are large rattlesnakes, abundant and typical of temperate zones of the southeastern region of the United States, running from southern Virginia to central Texas (Fig. 1A). They are semi-aquatic animals that live in humid areas, with fresh or brackish water sources, bodies of water, bays, sandbanks, lagoons, streams, ditches, and even at the bottom of rivers. They are known as cottonmouth because of the white color inside their mouths (Fig. 1B) [52–54].

Jia et al. (2008) built a cDNA library through an analysis of the RNAs transcribed in *A. p. leucostoma* venom gland, which has shown that PLA_2s are the proteins that constitute most of the venom of this species [2]. In fact, Lomonte et al. (2014) have also confirmed that PLA_2 is the main component of *A. p. leucostoma* venom through comparative proteomic studies [55]. Studies have shown that there are at least three isoforms of $sPLA_2$ in *A. p. leucostoma*: two basic (D49 and K49) and one acidic (D49) [41,56]. Even though the unusually high concentration of PLA_2 in the venom of *A. p. leucostoma* (approx. 35% m/m) has become a strong point of interest in research, the understanding of the pharmacological and toxic activities of these molecules in this specific venom remains limited [56].

Resende et al. (2017) isolated the acidic $sPLA_2$ (AplTx-I) from the venom of *A.p. leucostoma*, which corresponds to the prediction made by Jia et al. (2008) based in the analysis of the cDNA library of the venom gland of this species [41,56]. The biochemical, functional, and structural characterization of AplTx-I was carried out by Resende et al. (2017), showing that this enzyme has the structural characteristics expected for acid $sPLA_2$s present in snake venoms. However, contrary to other studies with acidic $sPLA_2$, the authors found that this enzyme does not have a high capacity for inducing edema, local and system myotoxicity, or blood clotting effects [56–59]. As others $sPLA_2$s from *A. p. leucostoma*, AplTx-I did not exhibit antibacterial activity [60]. Apart from the studies carried out by Resende et al. and Jia et al., there are no other studies on sPLA2 present in *A.p. leukostoma* venom [41,56,60].

In this work, we set to increase the knowledge on the basic sPLA2 (D49) (AplTx-II) from the *Agkistrodon piscivorus leucostoma* venom. Therefore, this enzyme was isolated to determine its structural, biological and toxic characteristics and kinetic parameters. The myotoxic and edematogenic properties produced by AplTX-II were also evaluated. We have further modeled the tridimensional structure of the holo form of the enzyme and docked the widely used synthetic chromogenic substrate, 4-nitro-3-octanoyloxybenzoic acid (NOBA) at the active site. Subsequent molecular dynamics (MD) simulations in explicit solvent allowed to construct a solvated, fully functional, pre-catalytic Michaelis Complex for this system.

2. Material and methods

2.1. Venom and reagents

Agkistrodon piscivorus leucostoma venom was provided by the National Natural Toxins Research Center (NNTRC) of the University Kingsville, Texas (USA). Reagents of sequencing or analytical grade were used.

2.2. Animals

Biological analyses were carried out with the authorization of the Committee for Ethics in Animal Experimentation of the Institute of Biology of UNICAMP, Protocol 3951–1. Swiss male mice (18–20 g), male Wistar rats (300–400 g) were purchased from the Laboratory Animals of the State University of Campinas (UNICAMP, Campinas, Brazil). The animals were maintained at 24–28 °C with food and water *ad libitum*.

2.3. Methodology

2.3.1. Purification of aplTX-II from Agkistrodon piscivorus leucostoma venom

Basic $sPLA_2$ (D49) from *Agkistrodon piscivorus leucostoma* venom was isolated in two chromatographic steps. The first step employed a gel filtration system with two Superdex-75 columns coupled in series. The chromatographic system was previously equilibrated with 50 mM AMBIC (ammonium bicarbonate, pH 8.0). Therefore, 20 mg of crude venom were dissolved in 200 μL of 50 mM AMBIC (pH 8.0), vortexed (2 min) and centrifuged 9000 rpm for 3 min, followed by the application of the supernatant on the system.

The elution was performed with 0.2 M AMBIC (pH 8.0), under a flow of 0.8 mL/min at 60 min, and the run was monitored at 280 nm. All the fractions collected were lyophilized and maintained at −30 °C. Two fractions (5 and 7) presented higher phospholipase activity. The fraction corresponding to Apl-7 is the acidic $sPLA_2$ (D49) previously described by our research group [56].

Reversed-phase high-performance liquid chromatography (RP-HPLC) used an HPLC-PDA 991 (Waters). The C18 column was a Discovery Bio Wide Pore (Supelco) (4.6 mm × 250 mm, 10 mm), previously equilibrated with buffer A (0.1% trifluoroacetic acid, pH 3.5).

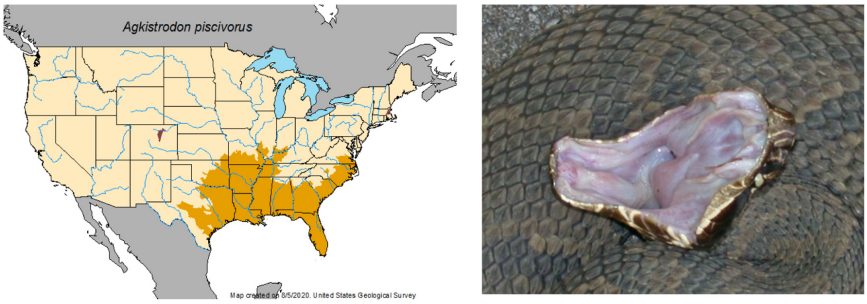


Fig. 1. A) Distribution map of *Agkistrodon piscivorus leucostoma* [121]. B) Detail of the color of the inside of the mouth (Photo: © 2007 Lisa Powers).

Next, 5 mg of the fraction 5 were dissolved in 120 μL 0.1% buffer A and 80 μL of 1 M AMBIC buffer (pH 8.0). The solution was vortexed (2 min) and clarified at 8000 rpm for 3 min. The pellet was discarded, and the supernatant was applied to the chromatographic system. The elution was monitored at 280 nm under a constant flow (1 mL/min) using a gradient (0–100%) of buffer B (66% acetonitrile in buffer A) [9,35,38,56,61]. To increase protein purity, the major fraction was re-chromatographed under the same conditions used in the first HPLC chromatographic step. The isolated protein was named ApITX-II.

2.4. Biochemical and structural characterization

2.4.1. Mono-dimensional electrophoresis

The homogeneity of the ApITX-II was evaluated by electrophoresis under reducing and non-reducing conditions, according to Laemmli, 1970 [62]. Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) was carried out discontinuously using a gel concentration of 5%–12.5%. Purified ApITX-II, as well as the molecular weight markers, were dissolved in sample buffer (0.075 M Tris-HCl, pH 6.8, 10% glycerol, 4% SDS, 0.001% bromophenol blue) and heated. The electrophoretic run was performed at 30 mA, and the gel was stained; the excess dye was removed using 7% acetic acid.

2.4.2. Two-dimensional electrophoresis (isoelectric focusing)

The theoretical pI of the purified toxin (ApITX-II) was calculated using a protein analysis tool (Compute pI/Mw tool), available at the Expasy server, and the experimental pI was determined according to the methodology described by the manufacturer (GE Healthcare Amersham Biosciences). The 2D electrophoresis consisted of two steps: isoelectric focusing and 1D SDS-PAGE. To begin with, ApITX-II was prepared in a rehydration solution (8 M urea, 2% CHAPS, 0.5/2% IPG buffer, 0.002% bromophenol blue, and 1 M DTT); this same solution was then incubated with a 7 cm strip (pH 3–10, linear) for 12–20 h. After rehydration, the strip was applied to an Ettan IPGphor 3 (GE Healthcare) isoelectric focusing system and later stored at −80 °C. Subsequently, the strip with ApITX-II was washed with DTT and iodoacetamide, each diluted in 5 mL of equilibration buffer solution (6 M urea, 2% SDS, 30% glycerol, 50 mM Tris- HCl, pH 7.4, 0.002% bromophenol blue). Then, the strip was applied to a 15% polyacrylamide gel. The gel was stained with Coomassie Blue G-250 and scanned in a GE Image Scanner III [56].

2.4.3. Enzymatic activity

The PLA_2 activity was assayed using the synthetic chromogenic substrate, 4-nitro-3-octanoyloxy benzoic acid (NOBA). All fractions obtained from the gel filtration and HPLC process were analyzed using this methodology. The colorimetric method used was described by Holzer and Mackessy (1996) and adapted for use in a 96-well microplate [63–66].

Different volumes of substrate and buffer were used to obtain NOBA concentrations between 0.125 and 40 uM. The standard assay mixture contained buffer (10 mM Tris–HCl, 10 mM $CaCl_2$ and 100 mM NaCl, pH 8.0), substrate (3 mM), 20 μL of water and 20 μL of PLA2 (1 mg/ml). The final volume of the assay was 260 μL. After adding PLA_2 (20 μg), the mixture was incubated for up to 40 min at 37 °C, with the reading of absorbance at intervals of 10 min. The enzyme activity, expressed as the initial velocity of the reaction (V_0), was calculated based on the increase of absorbance after 20 min.

The reaction product was colorimetrically quantified using a VERSA max microplate reader (Molecular Devices, Sunnyvale, CA) and calculated according to the change in absorbance. Mixture was incubated and monitored at 425 nm for 40 min at 37 °C, absorbance reading at intervals of 10 min. The PLA_2 activity was expressed as the initial rate of reaction (V_0) based on the increase in absorbance after 20 min and performed in triplicate.

2.4.4. Influence of substrate concentration, pH, temperature, and presence of divalent cations on the PLA_2 enzymatic activity

The influence of substrate concentration, pH, temperature, and presence of divalent cations on the PLA_2 used the same assay as described in subchapter 2.4.3, using the synthetic chromogenic substrate, NOBA (4-nitro-3-octanoyloxybenzoic acid) [56].

2.4.4.1. Influence of substrate concentration on the PLA_2 activity. The influence of substrate concentration on the PLA_2 activity of ApITX-II was assayed using different substrate concentrations NOBA (4-nitro-3-octanoyloxybenzoic acid): 0.125, 0.25, 0.5, 1, 2.5, 5, 10, 15, 20 and 40 uM.

2.4.4.2. Optimal pH for the enzymatic reaction. To evaluate the influence of the pH on the enzymatic assay we used the following buffers: sodium citrate (pH 4, 5 and 6), Tris-HCl (pH 7 and 8), and glycine (pH 9 and 10).

2.4.4.3. Optimal temperature for the enzymatic reaction. The optimal temperature for the enzymatic reaction was determined by enzymatic assay at the temperatures of 10, 15, 20, 25, 30, 35, 37, 40, 45, 50, 55 and 60 °C.

2.4.4.4. Influence of the presence of the divalent ions. The influence of 1 μM of several divalent ions (Ca^{2+}, Zn^{2+}, Cd^{2+}, Mg^{2+}, and Mn^{2+}) on the PLA_2 enzymatic activity was tested using buffers containing the following salts: $CaCl_2$, $ZnCl_2$, $CdCl_2$, $MgCl_2$ and $MnCl_2$.

2.4.5. Mass spectrometry analysis

2.4.5.1. Determination of ApITX-II mass. The isotope-averaged molecular mass of ApITX-II was determined using a NanoAcquity Ultra Performance LC, coupled with a nano-electrospray source, using a Waters Q-TOF Ultima mass spectrometer. 4.5 μg purified protein was injected into a C18 reverse-phase column UPLC (nanoACQUITY UPLC Waters) coupled with a mass spectrometry system. The elution used a flow rate of 600 nL/min (45 min) with a gradient of 0–50% acetonitrile in 0.1% formic acid. The instrument was operated in continuous mode, and MS data acquisition was carried out at 100–3000 m/z (speed of 1 s and delay of 0.1 s). The processing parameters were: output 6000–20,000 Da mass range with a resolution of 0.1 Da/channel [56].

2.4.5.2. Determination of the primary structure of ApITX-II

2.4.5.2.1. N-terminal sequencing. The N-terminal sequence of 50 μg ApITX-II was carried out in an automatic microsequencer PPSQ-33A (Shimadzu) according to the chemical process of sequencing by the N-terminal cleavage derivative method described by Edman (1949) [56,67].

2.4.5.2.2. ApITX-II digestion and analysis of digested proteins. Initially, 50 μg of ApITX-II and 10 μL of 50 mM ammonium bicarbonate (pH 8.5) were subjected to rapid centrifugation, followed by the addition of 5 μL dithiothreitol (DTT), vortexing and incubation at 56 °C for 25 min. Subsequently, 150 μL 5 mM iodoacetamide were added and the tube incubated at room temperature (30 min), protected from light. Again, 5 μL of DTT 5 mM were added and incubated (15 min) at room temperature, also protected from light.

After this preparation, ApITX-II was digested with a solution of 1 mM calcium chloride and 10 μL proteolytic enzyme (trypsin) and incubated for 16 h at 37 °C. After the digestion of the ApITX-II enzyme, the reaction was stopped with 100 μL of 0.4% trifluoroacetic acid (TFA), and the tube was vortexed and centrifuged (2500g, 10 min). The supernatant was transferred to a vial with 5 μL 1 N ammonium hydroxide.

The analysis of peptides obtained from the tryptic digestion of ApITX-II was performed at the National Bio-sciences Laboratory (LNBio) CNPEM-ABTLuS, Campinas, in a nanoACQUITY Ultra LC Performance coupled with a nano-electrospray source and Waters quadrupole-time-of-flight analyzer (Q-TOF) mass spectrometer Ultima (Milford, Massachusetts, USA). A study of overlapping peptides

obtained from the enzymatic digestion was performed. All fragments obtained from spectrometry were interpreted manually to derive *de novo* amino acid sequences. The alignment of the peptide fragments and the identification of homology with other PLA2s were carried out using available snake venom databases (NCBI-BLAST and UniProtKB) [56].

2.5. Pharmacological characterization

2.5.1. Local and systemic myotoxic activity

The ability of ApITX-II to induce local myotoxic and systemic effects was evaluated in groups of five mice (18–20 g), according to the methodology also used by Gutierrez et al. (2008). The animals received intramuscular (local myotoxicity) or intravenous injections (systemic myotoxicity) with different amounts of ApITX-II (5, 10, and 20 μg) dissolved in 50 μL of PBS buffer. Negative control was performed with mice receiving intramuscular or intravenous injections of 50 μL of PBS alone.

After ApITX-II administration, tail blood samples were collected into heparinized capillaries from the mice tails at various time intervals (0, 1, 3, 6, 9, 12, and 24 h). The blood samples collected were immediately centrifuged, and the plasma used for plasma creatine kinase (CK) quantification using a colorimetric assay, CK-NAC UV UniTest kit (Wiener lab). For this kinetic assay, 200 μL of substrate were pre-incubated for 4 min at 37 °C, followed by the addition of 5 μL of plasma.

The reaction production was monitored at 340 nm for 6 min, according to the manufacturer's instructions. The myotoxic activity was expressed in Units/Liter (U/L), in which 1 U is defined as the phosphorylation of 1 mmol of creatine/min at 25 °C.

2.5.2. Inflammatory activity (paw edema)

To evaluate the edema-forming activity caused by ApITX-II, groups of mice (18–20 g) received subcutaneous injections of different concentrations of toxin (5, 10, and 20 μg dissolved in 50 μL of PBS) in the right footpad. As a negative control, the left footpad received an identical injection of PBS (50 μL) alone. Footpad thickness was determined at different time intervals (0, 0.5, 1, 3, 6, 9, 12, and 24 h) after the injection of toxins and control solution, using a low-pressure spring caliper. The edema-forming activity was calculated as described by Lomonte et al. (1993), based on the percentage of increase in volume produced when the volume of the left and right footpads was compared.

2.5.3. Effect on coagulation

2.5.3.1. In vivo anticoagulant activity. Blood samples (10 mL) from male Wistar rats (300–400 g), obtained from the descending abdominal aorta, were added to 1 mL of 3.8% sodium citrate and centrifuged at 2500 ×*g* for 15 min (4 °C) to obtain platelet-rich plasma (PRP). Triplicate assays were carried out of the positive controls (20 U thrombin or 40 μg *B. jararaca* snake venom), negative control (saline), or with ApITX-II toxin (20–160 mg). Coagulant activity was estimated from the ability of the purified toxin to accelerate the coagulation of normal rat plasma and determined by the clotting time. If the sample did not induce plasma clotting within 10 min, it was considered to present no plasma coagulant activity.

2.5.3.2. In vitro anticoagulant activity. In vitro anticoagulant activity was determined according to the methodology used by Clement et al. (2012) with some adaptations. Briefly, different amounts of ApITX-II toxin (20–160 mg) in 0.15 M NaCl in a final volume of 100 ml were added to 0.4 ml of rat plasma with sodium citrate at 0.4%, which were incubated for 1 min at 37 °C. Subsequently, 100 μl of 50 mM $CaCl_2$ was added, and the clotting time was determined. Plasma samples treated in the same manner, but without the addition of acidic toxin, were used as control.

2.5.4. Antibacterial activity

The antibacterial potential of ApITX-II was assayed using the standardized single disk method in agar, as described by Bauer et al. (1966). For the assay, five bacterial strains were used: *Pseudomonas aeruginosa* ATCC 27853, *Pseudomonas aeruginosa* 31NM, *Escherichia coli* ATCC 5922, *Staphyloccocus aureus* BEC9393, and *Staphyloccocus aureus* Rib1. The isolated basic PLA_2 was tested in increasing concentrations up to 3 mg/mL. Commercially available antibiotics, Ofloxacin, and Imipinen were used as positive controls.

2.6. Modeling of the holo ApITx-II 3D structure in complex with NOBA

Currently, in the Protein Data Bank (PDB) [68], there is no record of X-Ray or NMR structures for the basic $sPLA_2$ (D49) protein from *A. p. leucostoma*, in neither apo nor holo states, hence the need for homology modeling. Therefore, we resorted to the Swiss Model server [69] and retrieved structures with the highest sequential identity as regards to the sequence of $sPLA_2$ D49 from *A. p. leucostoma* (UniProt AC: C0HKC3) [70]. We chose the template PDB ID: 1VAP, which presented the highest identity percentage, 99.19%, and 0.88 coverage (residues 17–139, as the initial 16 residues are a signal peptide).

According to UniProt [70], the sequence for the basic $sPLA_2$ (D49) *A. p. leucostoma* comprises a 17 residues signal peptide followed by 123 residues that make up the mature protein [41], and contains seven disulfide bonds that offer rigidity to the protein, in accordance to the expected number of disulfide bridges [34]. The chosen homology model checks these features, corroborating its status as a fit model.

Further refining of the homology structure included *i*) insertion of a Ca^{2+} ion, absent in the PDB ID:1VAP template structure, and *ii*) docking of the NOBA ligand, widely used in the enzymatic assays. Narrowing down our search in the PDB database, we identified a structure for porcine pancreatic IB PLA_2 (PDB ID:1FX9) [71], with 42.48% identity similarity, which presented both the Ca^{2+} ion as well as a ligand similar to NOBA. The 1FX9 structure was shown, by Pan et al. (2001) [71], to be in catalytic conformity, displaying a catalytic water bridging the Nδ atom of His48 and the oxygen of the sn-2 phosphate of the inhibitor MJ33 [71]. Alignment of the C_α atoms from the homology model and the aforementioned 1FX9 structure exhibited a root-mean-square deviation (RMSd) of 0.528 Å, demonstrating that the protein backbones are virtually identical.

While preparing our model, the active site residues were carefully checked regarding their protonation state. Considering past literature, as well as all the charged active-site amino acid residues in the immediate vicinity, we determined that the protonated nitrogen of H48 was indeed protonated at Nε, establishing a hydrogen bond with D90 [72]. All seven disulfide bridges were kept in the final structure.

Next, we placed NOBA (using Pymol's Pair Fitting function) [73] in the active site, following the position of the MJ33 inhibitor in the original 1FX9 template, as well as the proposed reaction mechanism [71]. Using Marvin Suite tools, we confirmed that the carboxylate and nitrite moieties of the NOBA ligand were most likely in a deprotonated state at pH = 7. Despite our best efforts, the hydrophobic region of the NOBA ligand clashed with residue K77. Subsequent classical molecular dynamics simulations solved the issue (see below).

We subjected our homology model, containing the coordinated Ca^{2+} and the NOBA ligand, to classical molecular dynamics (cMD). The goal was to *i*) stabilize the model and release critical contacts from the homology modeling; *ii*) determine the water occupancy at the active site; *iii*) observe the permanence of the NOBA ligand in the active site.

The parameters for NOBA were determined from an optimized structure at the HF/6-31G(d) level of theory in vacuum. The Antechamber tool, incorporated in the AmberTools of AMBER18 [74] was used with the GAFF force field [75] to derive the intramolecular and Lennard−Jones parameters, while the RESP (Restrained Electrostatic Potential) method [76] was used to derive atomic charges at the HF/6-31G(d) level of theory [77] to be consistent with the AMBER FF14SB force field used to parametrize the remaining atoms of the MM portion of the enzyme [78]. The system

was then solvated with TIP3 waters [79] in an octahedral box, whose faces were positioned at least at 12 Å from any protein atom. To neutralize the charge of the system, we added 7 Cl^- ions.

An energy minimization setup preceded the conventional molecular dynamics (cMD) simulations: in the first minimization cycle, only water molecules were minimized; in the second cycle, hydrogen atoms were minimized; in the third one, only the NOBA ligand, as well as the side-chain of the K77 residue, were minimized. Finally, in the fourth one, all side-chains were allowed to move while the backbone was kept under constraint, restrained with harmonic potential force constants of 50 kcal·mol^{-1} Å^{-2}. All bonds involving hydrogen atoms were constrained with the SHAKE [80] algorithm to use an integration step of 2 fs.

To perform cMD, the annealing of the system was conducted linearly for 80 ps from 0 to 310 K, at constant volume, while keeping the backbone of the protein constrained. Subsequently, we ran a 2 ns cMD within an NPT ensemble (at 1 bar and 310 K) with everything except water constrained to equilibrate solvent density, followed by 20 ns of unconstrained cMD. The NPT ensemble aimed to reproduce the conditions in which the enzymatic activity analysis of ApiTX-II was carried out.

2.7. Statistical analysis

Each experiment was repeated at least three times, and the results are reported as the means ± standard deviations (SD). The statistical

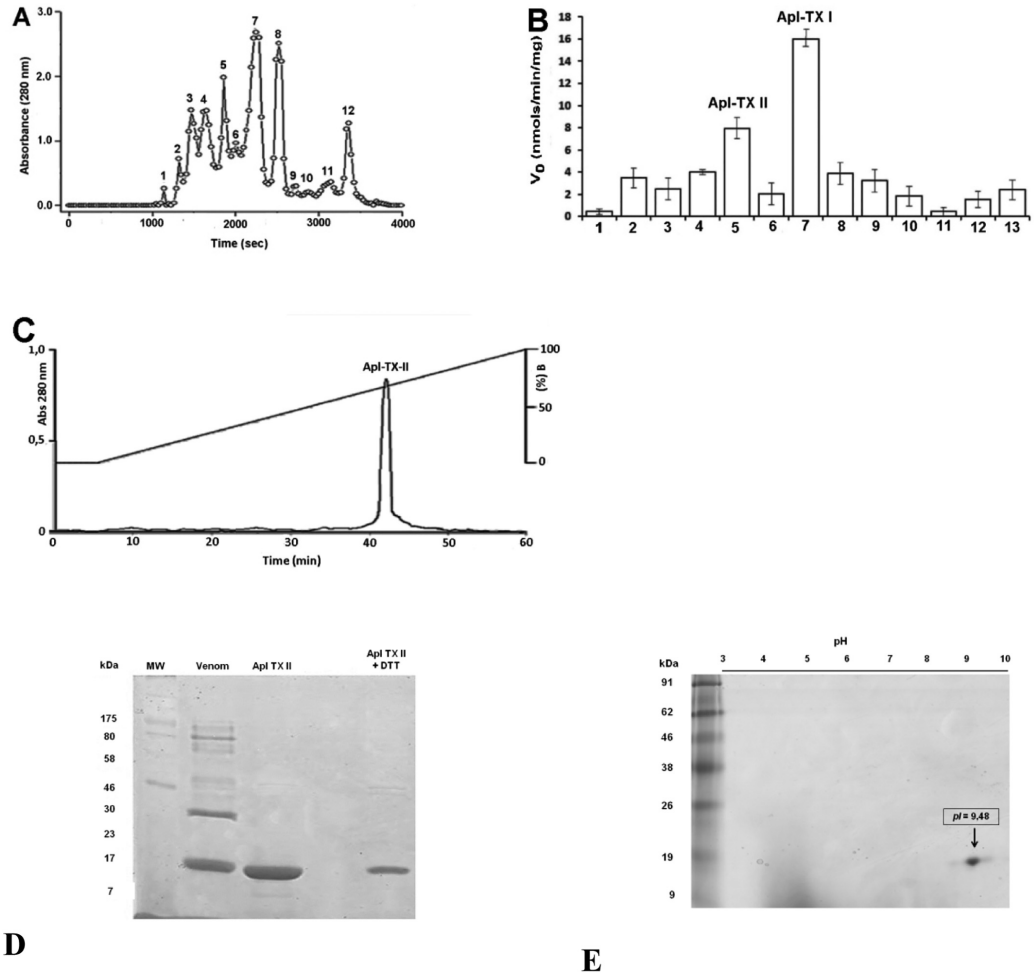


Fig. 2. Sequential purification of *A. p. leucostoma* venom. A) Fractionation of crude venom on gel filtration (Fraction 7 is ApiTX-I [56] and Fraction 5 is ApiTX-I); B) determination of the PLA_2 activities of the fractions obtained from gel filtration (fraction 7 corresponds to ApiTX-I); C) re-chromatography of fraction 5 (ApiTX-II). Profile elution of by C18 RP-HPLC (4.6 mm × 25 cm; 10 mm); D) electrophoretic analysis of ApiTx-II in SDS-PAGE under reducing and non-reducing conditions (SigmaMarker™ wide range, mol wt 6500–200,000 Da. Aprotinin, bovine lung (6500),α-Lactalbumin, bovine milk (14,200),trypsin inhibitor, soybean (20,000),trypsinogen, bovine pancreas (24,000), carbonic anhydrase, bovine erythrocytes (29,000), glyceraldehyde-3-phosphate dehydrogenase, rabbit muscle (36,000), ovalbumin, chicken egg (45,000), glutamic dehydrogenase, bovine liver (55,000), albumin, bovine serum (66,000),phosphorylase B, rabbit muscle (97,000),β-Galactosidase, *E. coli* (116,000) and myosin, porcine heart (200,000)); E) 2D electrophoresis (Isoelectric focusing and 1D SDS-PAGE).

significance of differences between the two groups was evaluated by Student's *t*-test, while the comparison of more than two experimental groups was determined by ANOVA tests, followed by Tukey's *Post Hoc* test using the Origin Pro 8 Software (OriginLab Corporation). Differences were considered as statistically significant if $p < 0.05$.

3. Results

3.1. Purification of AplTX-II from A. p. leucostoma venom

The whole venom of *A. p. leucostoma* was used to purify basic $sPLA_2$ D49 (AplTX-II) using two chromatographic methods (molecular exclusion with two Superdex-75 columns coupled in series and reverse-phase chromatography (Fig. 2). The fractionation profile obtained in a gel filtration system using the venom of *A. p. leucostoma* is shown in Fig. 2A. From gel filtration purification, we obtained 13 fractions, with the fraction 5 and 7 showing high PLA_2 activity (Fig. 2B). We identified fraction 7 as the AplTX-I [56]. Fraction 5 was identified as other $sPLA_2$ D49 and selected for this study.

Peak 5 was submitted to chromatography and re-chromatography (Fig. 2C) on C18 RP-HPLC. The purity and homogeneity of the fraction 5 protein obtained from the re-chromatography step were assayed by electrophoretic analysis in SDS-PAGE. Fig. 2D shows that the purified protein migrated as a single band of ≈14 kDa under non-reducing and reducing conditions, indicating its monomeric nature. The PLA_2 activity of the peak 5 protein was confirmed, and the enzyme was named AplTX-II. From 800 mg snake venom used, 15.8 mg of pure AplTX-II was obtained (≈ 2% m/m).

3.2. Structural characterization of AplTX-II PLA_2

Two micrograms of AplTX-II were used in order to analyze the electrophoretic mobility and relative molecular mass of AplTX-II assayed by SDS-PAGE, which was ≈ 14 kDa (Fig. 2D). This value was consistent with its isotope averaged molecular mass of 14,003.2256 Da, as determined by mass spectrometry (Fig. 3). One hundred micrograms was used to perform 2D electrophoresis. AplTX-II shows a theoretical pI of 8.96 and an experimental pI of 9.48 using 2D electrophoresis of the purified protein, confirming the basic nature of the purified molecule (Fig. 2E). After 2D electrophoresis, Melanie 2D gel analysis version 7.0 software was used to calculate pI experimental.

The primary structure of AplTX-II was obtained by N-terminal sequencing and digesting the toxin with different proteolytic enzymes and chemical reagents, together with analysis by tandem mass spectrometry and manual interpretation of mass spectrometry to derive *de novo* amino acid sequences (Tables 1 and 2). Ten peptides were identified (Table 1) from the digested protein, which allowed the classification of the protein as a $sPLA_2$ D49. A total of 95 amino acids were found, recovering about 77% of PLA_2 sequence isolated by Jia et al. (2008) [2].

The comparison of the primary structure of AplTX-II with other similar $sPLA_2$s present in snake venoms and deposited in the UniProt database [56] revealed that AplTX-II shares all the structural and functional motifs common to the basic $sPLA_2$s (amino-terminal region, site catalytic and Ca^{2+} loop) (Table 2). The presence of these structural motifs is in agreement with the enzymatic assay results, which demonstrated the capacity of AplTX-II to cleave a synthetic $sPLA_2$ substrate.

Given the high sequence identity between $sPLA_2$ from different species, it is possible to derive a high accuracy 3-D model of AplTx-II using homology modeling techniques. The 3-D structure of the toxin allows for a much more detailed understanding of the structure-function relationship and for structure-based inhibitor design.

A model was obtained using the X-ray structure with PDB ID:1VAP (UniProt Code P51972 – Table 2) as a template, which displayed a 99.19‰ identity percentage with the sequence of AplTx-II. At a second stage, we used the PDB ID:1FX9 structure as a template for Ca^{2+} insertion and substrate (NOBA) positioning in the active site. This structure presented a lower identity percentage

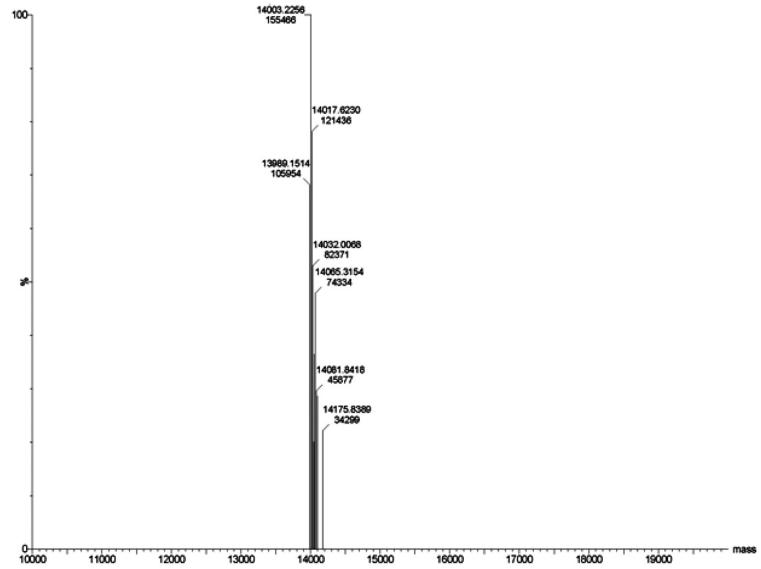


Fig. 3. Mass determination of the isolated basic PLA_2 (AplTX-II) by mass spectrometry. Deconvoluted spectrum, indicating the exact mass of AplTx-II of 14,003.2256 Da, confirming the relative molecular mass determined by SDS-PAGE (Fig. 2D).

Table 1
Peptide fragments resulting from the digestion of ApITx-II with proteolytic enzymes and a chemical reagent (CNBr), as identified by LC-MS/MS sequencing. All fragmentation spectra obtained were interpreted manually to derive *de novo* amino acid sequences.

Peptides	Monoisotops		
	Experimental mass (Da)	Theoretical mass (Da)	Accuracy
NLFQFEKLIKK (N-Terminal)	1406.8409	1406.8502	6.61
SGMLWYSAYGCYCGWGGQG (Ca^{2+} binding loop)	2426.0301	2426.0293	−3.30
DATDR	576.2403	576.2391	−2.08
CCFVHDCCYGK (HD Catalytic amino acids)	1276.4388	1276.4390	0.16
VTGCNPK	717.3377	717.3374	−4.18
CGGTNPCKK	906.3933	906.3939	0.67
QICECDR	865.3316	865.3312	−4.62
AAAICFRDNLK	1220.6231	1220.6234	0.25
TYDSK	612.2631	612.2638	1.14
TYWKYPK	984.4941	984.4959	1.83

(42.48%), hence not being chosen in the first place as a main homology modeling template. The derived model was verified through visual inspection, and checked with the QMEAN and ProSAII [81] scoring functions. The ProSAII score for the 1VAP X-ray template structure was −4.87, while our model scored −4.67 and, once minimized, −4.37. The score falls well within the typical values for proteins of the same size, as can be seen in Fig. 4A and B.

This indicates that the structure of our model is native-like. The same can be said about the QMEAN score, which was 0.59, again well within the typical values for proteins of the same size (Fig. 4A and B). The QMEAN score of the model falls within less than one standard deviation from the average QMEAN of proteins of the same size, further confirming the reliability of the derived model.

Once assembled, we superimposed the model with the four X-ray structures with the highest identity percentage available in Swiss Model (Fig. 5A and B). We also aligned the model with five other PLA_2 X-ray structures having co-crystallized inhibitors (Fig. 5C). Finally, we attested the integrity of the model through cMD simulations. We performed a 22 ns cMD simulation that allowed to identify the main interactions between NOBA and the active site residues, as well as those with the proposed catalytic water and Ca^{2+} (Fig. 4).

The conserved architecture of the $sPLA_2$ active site is characterized by the H48/D99/Calcium motif. Previous proposals for the $sPLA_2$ mechanism suggest the formation of the tetrahedral intermediate by the nucleophilic attack of a water molecule, which is activated by the H48 [82]. During the cMD simulation, the conformational arrangement of the active site residues was kept perfect for catalysis: the postulated catalytic water molecule was coordinating the Ca^{2+} ion (which lowers its pKa), and another water molecule was establishing hydrogen bonds with both the catalytic His48 (*i.e.* well positioned for deprotonation) and D90, keeping proximity (3.4 Å) to the ester carbon of the NOBA ligand that makes the sn2-acyl bond (Fig. 6A and B).

3.3. The enzymatic activity of AplTX-II

We used the chromogenic substrate 4-nitro-3-(octanoyloxy) benzoic acid (NOBA) to perform all assays of phospholipase A_2 activity of AplTX-II (Fig. 7). The analysis of the effect of substrate concentration on the AplTX-II activity is presented in Fig. 7A. AplTX-II showed a Michaelian behavior when the substrate concentration ranged between 0 and 40 mM, and kinetic parameters, K_M and V_{max}, were calculated at 4.82 μM and 17.43 nmol/min/mg, respectively (Fig. 7B). The optimum temperature and pH were determined at 37 °C (Fig. 7C) and pH 8.0 (Fig. 7D). As others $sPLA_2$, AplTX-II was strictly dependent on the Ca^{2+} cation concentration for its enzymatic activity. When Ca^{2+} was substituted by 10 μM of other divalent cations (Cd^{2+}, Mg^{2+}, Mn^{2+}, Zn^{2+}), the AplTX-II activity decreased significantly. Also, when we mixed these other positive ions with 10 μM of Ca^{2+}, the enzyme activity was not recovered (Fig. 7E).

3.4. Pharmacological characterization of AplTX-II

In vivo, *ex vivo* and *in vitro* studies of AplTX-II systemic and local myotoxicity, edema induced, effects on plasma coagulation, and antibacterial potential were carried out to understand the functional roles and molecular and pathophysiological mechanism of action of envenoming by AplTx-II.

3.4.1. Systemic and local myotoxicity

Basic $sPLA_2$s (D49) are generally myotoxic. This was also the case with AplTx-II, which was capable of inducing local and systemic myotoxicity. Local myotoxicity was observed when the animals received intramuscular injections of AplTX-II. A dose-dependent elevation of the plasma creatine kinase activity was observed, reaching maximum levels after 1.5 h of injection. The plasma creatine kinase activity returned to the basal level after 24 h (Fig. 8A). AplTX-II was also capable of inducing systemic myotoxicity upon intravenous injection. A dose-dependent response in the plasma creatine kinase activity was observed and reached a maximum after 1 h. The plasma creatine kinase activity did not return to the basal level 24 h after injections of 10 and 20 μg of the enzyme (Fig. 8B).

Table 2
Sequence alignment of AplTx-II with other similar snake venom basic $sPLA_2$ (D49). Protein codes in the UniProt database are shown on the left of each primary structure The N terminal sequence is shown in a green box. The Ca^{2+} binding loop is shown in a gray box. Catalytic amino acids are shown in a yellow box. ***ApITX-I:** acid PLA_2 D49 from *Agkistrodon piscivorus leucostoma venom* [56]; **C0HKC3**: *Agkistrodon piscivorus leucostoma* [2]; **P51972**: *Agkistrodon piscivorus piscivorus* [41]; **C0HKC4**: *Agkistrodon piscivorus conanti* [41]; **C0HLF0** *Porthidium ophryomegas* [122]; **P04417**: *Agkistrodon halys blomhoffii* [123].

UniProt Code *	Sequence
ApITX-II	NL F QFEKL IKK_____SGM LWYSAYGCYCGWGGQGRP KDAT DRCCFVHDCC
ApITX-I	DLMQFET L IMK I A KRSGM FWYSAYGCYCGWGGQGRPQDAT DRCCFVHDCC
C0HKC3	MRTLWIMAVLLLGVEGNL F QFEK L IK KMTGKSGM LW YSAYGCYCGWGGQGRP KDAT DRCCFVHDCC
P51972	NL F QFEK L IK KMTGKSGM LW YSAYGCYCGWGGQGRP KDAT DRCCFVHDCC
C0HKC4	NL F QFEK L IK KMTGKSGM LW YSAYGCYCGWGGQGRP KDAT DRCCFVHDCC
C0HLF0	NL F QFRKMIK KMTK KE P V VY YAFYGCYCG K GGRGKP KDAT DRCCFVHDCC
P04417	HL L QFRKMIK KMTG KE P V I S YAFYGCYCG SGGRGKP KDAT DRCCFVHDCC
ApITX-II	YGKVTGCNPK CGG TNP CKKQ I CECDR AAAICFRDNL KTYDS KTYWK YPK
ApITX-I	YGKVTGCDPK L DSYTY SVENGD VVCGGNDP CKK E I CECDR AAAICFRDNKVTYDNKYWRFPPQNCKEESEPC
C0HKC3	YGKVTGCNPK MD IYTY SVENGN IVCGGT NP CKKQ I CECDR AAAICFRDNL KTYDSK TYWK YPKKNCKEES EPC
P51972	YGKVTGCNPK MD IYTY SVDNGN IVCGGTNP CKKQ I CECDR AAAICFRDNL KTYDSK TYWK YPKKNC KEES EPC
C0HKC4	YGKVTGCNPK MD IYTY SVDNGN IVCGGTNP CKKQ I CECDR AAAICFRDNL KTYDSK TYWK YPKKNC KEES EPC
C0HLF0	YEKVTGCNPK WGYYTYSMN KQ IVCGGDDP CKKQVCECDK AAAICFRDNL KTY K K KYMS FPNF FC TDPSEKC
P04417	YEKVTGCKPKWDDYTYSWK NGDIVCGGDDP CKK E I CECDR AAAICFRDNL KTY K K RYMA YPD I LC SSKS EKC

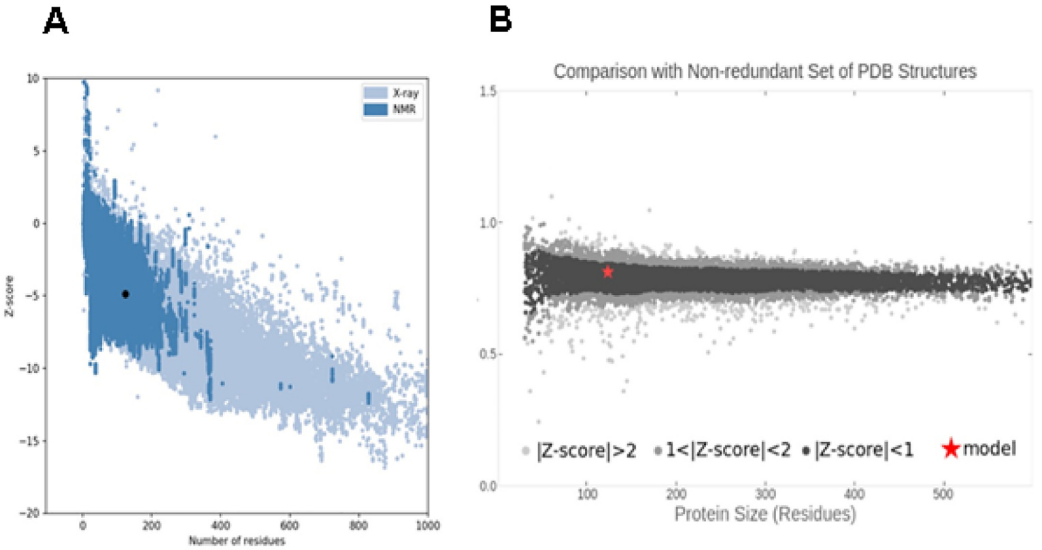


Fig. 4. A) ProSall Scoring Plot for the minimized state of the homology model created using 1VAP x-ray structure as a template. B) ProSAII scoring is −4.37, once minimized. The score falls well within the typical values for proteins of the same size.

3.4.2. Edema

The subplantar injection of AplTx-II induced local edema in the mice footpad assay (Fig. 8C). The edema-forming activity peaked after 30 min, achieving the highest value of ≈ 50% upon injection of 20 μg of AplTx-II.

3.4.3. Anticoagulant effect

AplTx-II was not able to produce any degree of blood clotting when compared to the negative control for 10 min. Similarly, AplTx-II also did not cause coagulation in rat plasma (*in vitro*), up to a concentration of 160 μg/mL. The mean plasma clotting time of rats incubated with purified AplTx-II was similar to the mean clotting time of the control group (data not shown).

3.4.4. Antibacterial effect

The AplTx-II antimicrobial potential to inhibit the growth of Gram-negative and Gram-positive bacteria was assayed by the agar diffusion method. AplTx-II was unable to mediate an antibacterial effect against the bacteria tested (*Pseudomonas aeruginosa* ATCC 27853, *Pseudomonas aeruginosa* 31NM, *Escherichia coli* ATCC 25922, *Staphyloccocus aureus* BEC9393 and *Staphyloccocus aureus* Rib1), up to concentrations of 3 μg/mL (data not shown).

4. Discussion

Snake venoms are characterized as materials with a complex combination of bioactive proteins and peptides. The toxic molecules present in

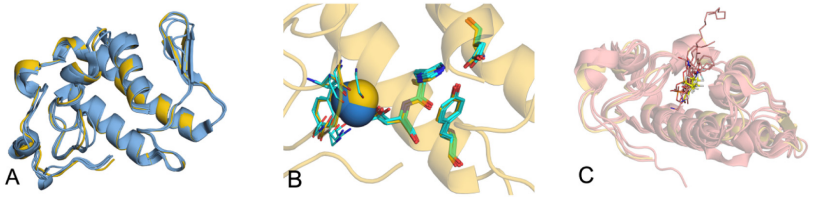


Fig. 5. A) AplTx-II homology model after cMD (yellow cartoon) superimposed with the four X-ray structures with the highest sequence identity (blue cartoon). The four structures have PDB IDs: 1VAP (Structure of Basic Phospholipase A_2 from *Agkistrodon piscivorus piscivorus*, used as a template), 1JIA (Structure of Basic Phospholipase A_2 from *Agkistrodon halys Pallas*), 1PSJ (Structure of Acidic Phospholipase A_2 from *Agkistrodon halys pallas*), and 1C1J (Structure of Cadmium-substituted Phospholipase A_2 from *Agkistrodon halys Pallas*). The superposition of structures shows the conformational similarity between the proteins. B) Active site residues and Ca^{2+} coordinating residues in the loop from both AplTx-II homology model (protein structure in the yellow cartoon; residues in yellow sticks, Ca^{2+} in the yellow sphere) and the four x-ray structures with the highest sequence identity (residues in cyan sticks, Ca^{2+}-only present in the structures 1psj and 1c1j - represented as blue spheres). Residue alignment shows high conformational similarity and consistency in the orientation of the active site residues. C) Alignment of the homology model (yellow cartoon and NOBA ligand in yellow sticks) with theX-ray structures of PDB IDs 1OXL (Structure of Phospholipase A2 from *Daboia russelli pulchella*), 1FX9 (Structure of Dimeric Phospholipase A2 from *Sus scrofa* with MJ33 Inhibitor), 1POE (Structures of Free and Inhibited Human Secretory Phospholipase A2 from Inflammatory Exudate), 1FV0 (Structure of a Complex formed between Phospholipase A_2 and Aristolochic Acid), and 5G3N (Structure of a Novel Secreted Phospholipase A_2 Inhibitor Azd2716), all of which contain a PLA_2 inhibitor (pink cartoon, pink sticks). Visual inspection shows high consistency in the positioning of the inhibitors and the modeled NOBA substrate in the active site.

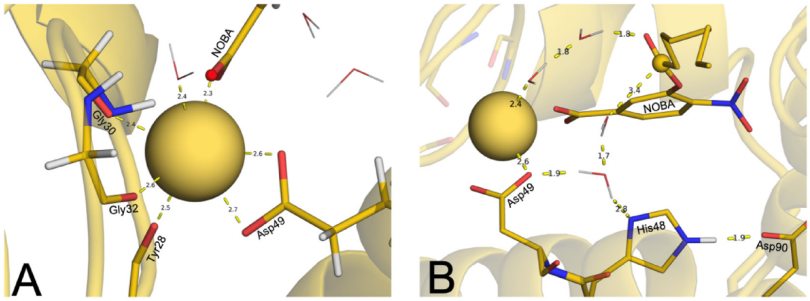


Fig. 6. A) The Ca^{2+} ion and corresponding loop of the homology model created for Apl-TxII. The calcium ion coordinates six oxygen atoms, including one from the substrate NOBA (represented as a red sphere), and one from a water molecule. B) NOBA bound to the active site, in close proximity to the Ca^{2+} ion. The distances between NOBA, the Ca^{2+} ion, the key active site residues (His47, Asp49, Asp90) and water molecules are shown. The structure gives support to the previous proposal for the reaction mechanism of this enzyme.

this material are selective and specific to their pharmacological targets and are dependent on the synergic action of several components [83–85]. The integrative action between the toxic components of the venoms, as phospholipases A_2, L-amino acid oxidases, metalloproteases, or serine proteases can disrupt the human hemostatic system through different mechanisms [9,15,26,30,34,53].

The analysis and understanding of the action of the venom's components are essential to help decreasing their biochemical, physiological, and pathological toxic effects. Moreover, these studies can be used for the development of new biotechnological and pharmaceutical products. Several components of the snake venoms as peptides and small proteins have shown a wide range of pharmacological activities, such as antimicrobial, antihypertensive, analgesic, antitumor, among others [9,16,25,32,33,35,42,48,56,86–88].

In this work a basic PLA_2 has been effectively purified by chromatographic techniques (Fig. 2A–C) [38,43,61,63,64,89]. Basic $sPLA_2$ (D49) isolated from venom of the snake *A. p. leucostoma* in this work was named as AplTX-II. This toxin was isolated, purified, and characterized. We determined the pharmacological and structural profiles of this new basic $sPLA_2$, which is within those molecules that present great interest by their toxic versatility and biotechnological potential [32,86]. Upon snake bite and envenoming, these enzymes develop important myotoxics and edematogenics effects on the organism of the prey or victim. Due to these critical toxic effects, research has been conducted on inhibitors, synthetics or naturals, of the $sPLA_2$ that may be used as an adjuvant to the treatment with antiophidic serums [33,90].

Through SDS-PAGE, AplTX-II showed an apparent MM of ≈ 14 kDa, both in conditions reducing, as well as non-reducing. This behavior of AplTX-II indicates its monomeric nature common to other snake venom $sPLA_2$s that also consist of a single polypeptide chain (Figs. 2D and 3) [57,83,91,92]. In addition, partial primary structure of AplTX-II was determined by sequencing the N-terminal region and digesting the enzyme with different proteolytic enzymes and chemical reagents.

The fragments obtained from digestion were analyzed by tandem mass spectrometry and manual interpretation of mass spectrometry to derive *de novo* amino acid sequences. Analyzes showed that primary sequence of AplTX-II determined a total of 95 amino acids that cover 77% of the sequence identified by Jia et al. (2008) [2] through the proteomic analysis of *A. p. leukostoma*. Analyzing the results, AplTX-II is probably the same basic $sPLA_2$ (D49) identified by Jia et al. (2008) [2] in proteomic studies of *A. p. leukostoma* (Tables 1 and 2).

When analyzing the fragments obtained from mass spectrometry, it appears that the primary structure of AplTX-II had the same structural motifs as other basic PLA_2s derived from snake venoms and that are deposited in the UniProt database [93]. The structural motifs common to all basic PLA_2s have: i) amino-terminal region (NLFQFEKLIKK); ii) catalytic site (CCFVHDCCYGK), and iii) Ca^{2+} coordinating loop (CGWGGQG) (Table 1). The presence of these conserved motifs is consistent with the results of the enzymatic assay where AplTX-II basics demonstrated the ability to cleave the synthetic substrate (NOBA) [63] (Table 2).

Although we have not been able to find all cysteine residues apparently, similarly to other basic $sPLA_2$s, AplTX-II also has 14 cysteines and 7 disulfide bonds, crucial structural characteristics of PLA_2s to stabilize the three-dimensional structure (Table 2). Additionally, by analyzing the obtained amino acid fragments it can be seen that AplTX-II maintains a high degree of structural similarity with other basic $sPLA_2$s. These enzymes have a large amount of hydrophobic and basic amino acids [63].

There is evidence that the high amounts of hydrophobic and basic amino acids in basic $sPLA_2$ are likely to mediate their toxic effects [35,38]. The pharmacological sites where basic phospholipases act must be complementary to each other in terms of weak interactions such as hydrophobicity, charges and van der Waals [31,56]. Studies have suggested that the distribution of hydrophobic and positively charged amino acids on the surface of $sPLA_2$s play an essential role mainly in the myotoxic activity presented by basic PLA_2 [38,43,44].

We proceeded by homology modeling the 3D structure of AplTx-II. As discussed before, the high quality of the templates and the results of all evaluation parameters show that the 3D model is highly reliable. Superposition of the modeled structure with the four most closely related $sPLA_2$s with known X-ray structure (Fig. 3A) shows a striking similarity, indicating that the fold is very conserved and highly plausible. At the active site and Ca^{2+} coordinating loop, all residues show an evident overlap, consistent with the aforementioned high level of conservation between organisms (Figs. 4 and 5) [64].

We also modeled the artificial substrate NOBA at the active site and the water solvent around the system, and ran cMD simulations of the system. The alignment of this model with five PLA_2 X-ray structures that possess co-crystallized inhibitors (Fig. 6), reveals a high similarity on the pose among NOBA and all of them.

Maybe the most relevant is that the arrangement of the substrate, active site and nearby water molecules (Fig. 6) is very favourable for the reaction to take place, according to the classical description of the $sPLA_2$ mechanism. However, the modeling was not guided to achieve such structural arrangement. This indicates that the NOBA artificial substrate will react through the same mechanism of the phospholipidic physiologic substrates, justifying its extensive use as a model substrate to study the physiologic activity of $sPLA_2$. Further computational studies addressing the characterization of the catalytic mechanism of NOBA and phospholipid substrates with atomic resolution would be beneficial to solve this question definitively.

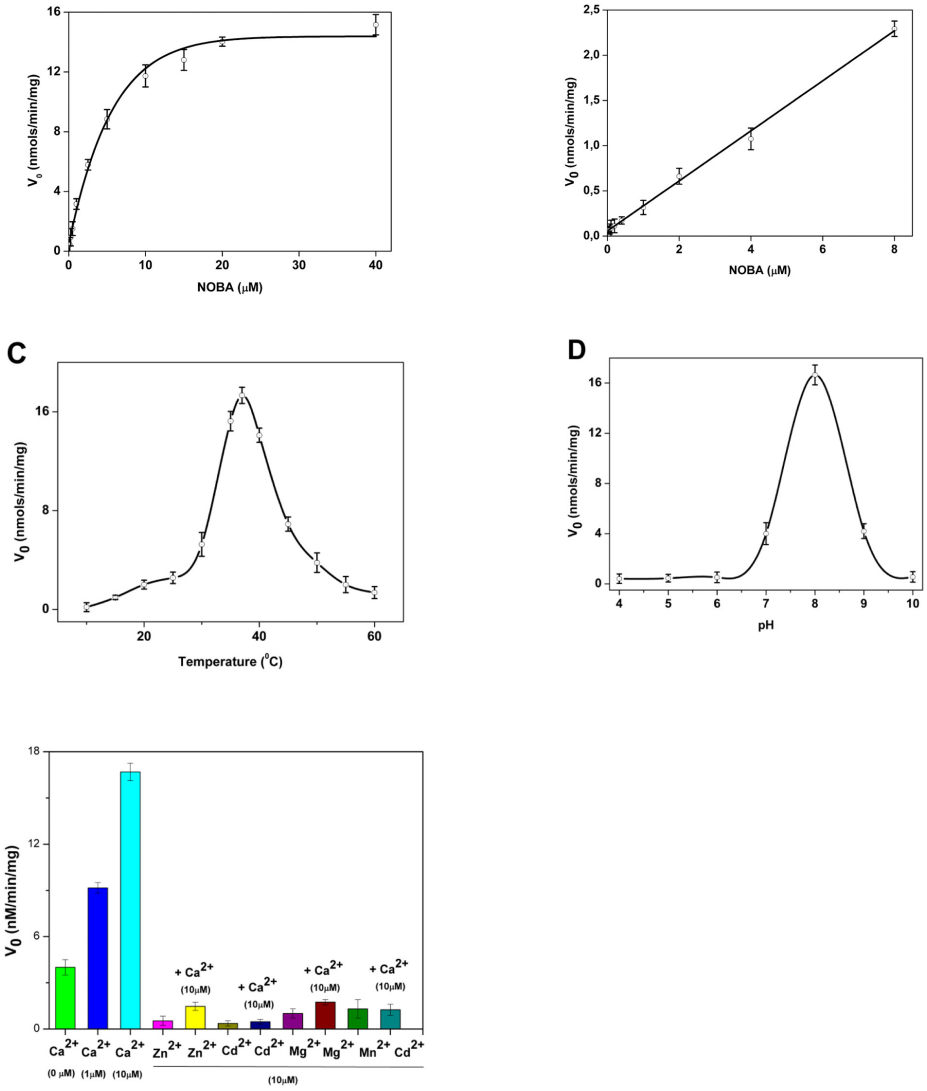


Fig. 7. ApITx-II phospholipase A_2 enzymatic activity. A) Effect of the NOBA substrate concentration on the initial reaction rate, V_0; B) K_M (4.82 μM) and V_{max} (17.43 nmol/min/mg) obtained from a Lineweaver-Burk plot; C) the effect of the temperature on the PLA2 activity; D) the effect of the pH dependence on the PLA2 activity; E) effect of a set of divalent cations on the PLA_2 activity.

The kinetic studies used the soluble synthetic chromogenic substrate NOBA [38,63]. The values of Km and Vmax were 4.82 mM and 17.43 nmol/min/mg. The maximum enzymatic activity occurred at ≈ 37 °C and pH ≈ 8 (Fig. 5B–D). These results are similar to those of other basic $sPLA_2$ when using the same substrate [38,92]. In addition, Jia et al. (2017) [41] also observed phospholipase activity against micellar L-a-phosphatidylcholine using this enzyme and a colorimetric assay. These two enzymatic assays confirm that the catalytic site has an H48 and a D49, essential to the catalytic activity. The homology model built here further places H48 and D49 in a perfect position for catalysis (Fig. 7A–D).

Basic $sPLA_2$ (D49) is highly dependent on the presence and concentration of Ca^{2+} [31,63]. The effect of the divalent calcium cation on the enzymatic activity was analyzed, resulting in that ApITX-II was dependent of Ca^{2+}. Indeed, in the absence of Ca^{2+}, ApITX-II showed a significant reduction of activity [92,94]. However, when Ca^{2+} was added to the reactional mixture at concentrations of 1 μM and 10 μM, ApITX-II recovered its catalytic activity (Fig. 4E). If Ca^{2+} was absent and other

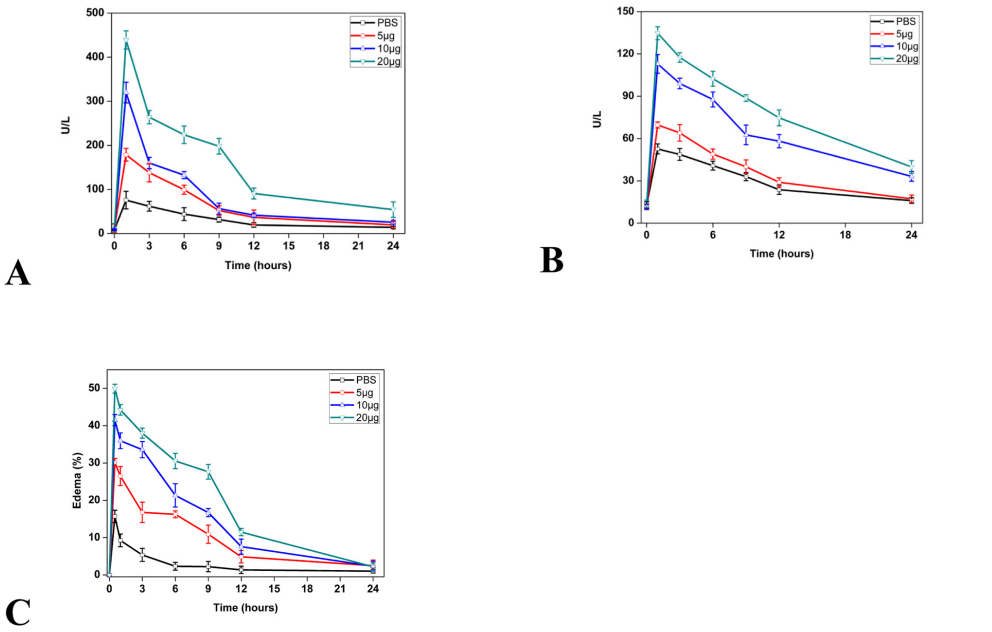


Fig. 8. AplTx-II induced local (A) and systemic (B) myotoxic activity as measured by the plasma creatine kinase activity. (C) Dose-dependent mice paw edema provoked by AplTx-II.

divalent cations (10 μM of Cd^{2+}, Mg^{2+}, Mn^{2+}, or Zn^{2+}) were added to the reaction mixture, the enzymatic activity was drastically reduced. When Ca^{2+} (10 μM) was added, *a posteriori*, to the reaction mixture that contained other divalent cations, the enzymatic activity remained diminished (Fig. 7E) [95].

There is already agreement among researchers that in basic $sPLA_2$, the Ca^{2+} ion is crucial for the electrophilic behavior of the active site and for the stabilization of the flexible calcium-binding loop responsible for the optimization of the conformation of the chemical reaction transition state. It seems that other divalent cations (Cd^{2+}, Mg^{2+}, Mn^{2+}, Zn^{2+}) do not allow the formation of the optimal interactions to stabilize the transition state. These divalent cations have inhibited the calcium-mediated activity by competitive processes, and seem to be unable to provide a suitable catalytic environment to position the substrate for catalytic action, probably due to their sizes and electronic structures [38,43,63]. In fact, the Ca^{2+} is smaller than the other transition metals (Zn^{2+}, Cd^{2+}, Mn^{2+}), making it a harder cation; Ca^{2+} also lacks the *d* orbitals that all these transition metals possess, thus making more ionic/less covalent interactions. Ca^{2+} is significantly larger than the much harder Mg^{2+} ion. However, the tightness with which the latter coordinates water molecules might place additional difficulties for NOBA to coordinate the Mg^{2+} ion, as it will need to displace the tightly-bound water molecules first.

It is interesting to note that the addition of Ca^{2+} *a posteriori* was unable to recover the full enzyme activity. This indicates that the enzyme has a higher affinity for Zn^{2+}, Cd^{2+}, Mn^{2+} or Mg^{2+} than to Ca^{2+}. Affinity and reactivity follow opposite trends for these divalent cations. Further computational studies using quantum mechanical methods would be needed to definitely elucidate the origin of this fascinating behavior.

Earlier evidence has strongly shown that basic $sPLA_2$s are among the major mediators of myonecrosis, hemolysis, mast cell degranulation, and edema formation [24,82–85,96–102]. In this context, the pharmacological activities of AplTX-II [103] local and systemic damage to skeletal muscle tissue and induction of paw edema, effect on the coagulation (*in vivo* and *in vitro*), and antibacterial activity were investigated here.

Like other basic $sPLA_2$ (D49), AplTX-II has shown both local and systemic myotoxic activities. We observed that both local and systemic myonecrosis caused by AplTX-II is both concentration- and time-dependent (Fig. 8A and B). This characteristic is associated with its high selectivity for, probably, myotoxic [17,97–99,104,105] specific targets. It is possible that the basic $sPLA_2$ toxicity towards muscle cells is induced through enzyme sites distinct from the catalytic site for membrane phospholipids [35,38,44,61,96].

AplTX-II showed greater activity of creatine kinase (CK) when injected directly into muscle tissue (450 U/l, ≈ 1 h) compared intravenous injection, where the activity of CK was significantly lower (130 U/l, ≈ 1 h). The differences observed may be related to the mechanism of accessibility of AplTX-II to muscles (intramuscular injection) and sequestration of toxins by binding to low-affinity sites in the membranes of other types of cells (intravenous injection).

The molecular mechanism of the myotoxicity produced by snake venom' $sPLA_2$s is still not fully understood. Some studies suggested that it is dependent on phospholipid hydrolysis. However, the effect is also related to the precise site of binding of $sPLA_2$ to cell receptors, which plays a decisive role in the location, extent of damage at the muscular tissue. It has been proposed that $sPLA_2$s may induce muscle cell damage by affecting the integrity of plasma membranes, but also by causing hypercontraction and other intracellular effects. Studies have shown that acidic $sPLA_2$ (D49), although enzymatically active, have much lower myotoxicity than basic $sPLA_2$ (D49). Even more impressive are the $sPLA_2$ (K49) isoforms, which do not have enzymatic activity, but are still capable of causing severe damage to muscle cells [17].

Earlier studies on the different myonecrotic capacity of the (D49) and (K49) $sPLA_2$s indicated that the difference might arise from their C-terminal regions, in particular the 115–129 region, where the two isoforms present significative sequence divergence. These findings support

the suggestion that the (K49) and (D49) isoforms produce myotoxicity through distinct molecular mechanism(s) [17,106–109].

The results derived from this work are in agreement with the proposed explanatory model for the pharmacological action of snake venom PLA_2. An important difference between the pharmacological profile of the *Bothrops* and *Crotalus* venoms' $sPLA_2$s is that the first induces local muscle damage, focused at the injection site, and the second induces systemic myotoxicity. Systemic envenoming derived from snake bites of the *Crotalus* genus causes a significant increase in plasma CK activity, breakdown of muscle fibers, and myoglobinuria. The results found here show that the myotoxic profile of ApITX-II of *A. p. leucostoma* (*Agkistrodon* genus) shares the myotoxic profile of the *Bothrops* genus, and contrasts with the one from the *Crotalus* genus.

The results obtained here also indicate that ApITX-II induces a concentration- and time-dependent inflammatory response. The edematous response started right after subcutaneous injection in the footpad, reaching a maximum value at 30 min (50% induction of edema). In fact, the production of local inflammation has often been seen as a typical pathological effect in viperid snake envenomation [4,7,40,52,83,110–112]. Some biochemical mechanisms can explain the induced paw edema: i) phospholipid hydrolysis with subsequent release of biologically active fatty acids and platelet-activating factor; ii) mast cell histamine release; iii) interaction of $sPLA_2$ with cell receptors, through regions different from the catalytic site and independent from catalytic action (Fig. 8C) [50].

$sPLA_2$ is a key enzyme in the formation of lipid mediators of inflammation processes, such as thromboxane, leukotrienes, and prostaglandins from arachidonic acid. However, several studies indicate that the inflammatory effect derived from snakes can be very complex because, in addition to the immediate effects at the bite site, pro-inflammatory chemical messengers (interleukin-1, interleukin-6, and interleukin-8) are also activated, which can also be associated with the inflammatory pathophysiological effects [50,84,85,88,94].

The possible role of the pro-inflammatory messengers in snakebite inflammation is reinforced by studies that indicate that the capacity of basic $sPLA_2$s (D49) to induce inflammation is dependent on the presence of the Ca^{2+} ion, showing that catalytic action is needed to induce edema. However, other studies also show that the catalytically-inactive $sPLA_2$s (K49) also induce the formation of edema, in similar levels with $sPLA_2$ (D49), suggesting that multiple different mechanisms are behind edema formation [43,44,89,113–120].

Studies suggest that the edematogenic effect of $sPLA_2$ is partially dependent on the extent of positive charges in the primary structure (this is also known to be the case with myonecrotic effects). Hydrophobic residues also increase the edematogenic effects. ApITx-II shows a high content of basic (positive) and hydrophobic residues, which is in line with the edematogenic effects observed here.

ApITX-II did not show any effect on the coagulation, *in vivo* or *in vitro*, up to a concentration of 160 μg/ml. The action of the ApITX-II on the blood clotting was similar to the negative control for 10 min.

ApITx-II also did not show antimicrobial potential, in terms of the inhibition of the growth of Gram-negative and Gram-positive bacteria (*Pseudomonas aeruginosa* ATCC 27853, *Pseudomonas aeruginosa* 31NM, *Escherichia coli* ATCC 25922, *Staphyloccocus aureus* BEC9393 and *Staphyloccocus aureus* Rib1), up to concentrations of 3 mg/ml. In effect, many studies have shown that the significative antibacterial and myotoxic effects of $sPLA_2$ are related to the C-terminal region. As discussed above, this region is dissimilar in sequence between (D-14) and (K-49) $sPLA_2$, the latter possessing antimicrobial activities [30,35,38]. ApITX-II and the basic $sPLA_2$ identified by Jia et al. [2,41,60] are most probably the same enzyme. In a previous work, Jia et al. (2018) [60] also concluded that this enzyme does not present activity antibacterial against gram-positive and gram-negative bacteria (*Bacillus subtilis*, *Staphylococcus aureus*, *Escherichia coli*, *Vibrio cholerae*), which is consistent with our results.

5. Conclusion

This work increases the understanding of the characteristics of a basic $sPLA_2$ present in the *A. p. leucostoma* venom, named here as ApITx-II. This enzyme produces significant toxic effects that can cause severe complications in cases of snakebite accidents. Based on the findings in this work, ApITx-II induces pathophysiological effects similar to other basic $sPLA_2$, in particular, those isolated from *Bothropics* venoms. Due to the many interesting pharmacological and biotechnological properties of phospholipases of snake venoms, it is vital broadening the studies on these (and other) enzymes, in particular those present in the *A. p. leukostoma* venom, which is a material that remains understudied and poorly understood.

CRediT authorship contribution statement

Letícia M. Resende: Formal analysis, Investigation, Methodology. **José R. Almeida:** Formal analysis, Investigation, Methodology. **Tatiana A. Guaraca-Medina:** Investigation, Methodology. **Matilde F. Viegas:** Investigation, Methodology, Writing – original draft. **Andreimar M. Soares:** Investigation, Methodology. **Maria J. Ramos:** Conceptualization, Writing – original draft, Writing – review & editing. **Pedro A. Fernandes:** Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing. **Sergio Marangoni:** Conceptualization, Formal analysis, Project administration. **Saulo L. Da Silva:** Conceptualization, Formal analysis, Project administration, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by UID/MULTI/04378/2019 and UIBD/50006/2020 with funding from FCT/MCTES through Portugal's national funds. We further thank the FCT for MFV's PhD grant (SFRH/BD/131968/2017) as well as for funding project PTDC/QUI-OUT/1401/2020. Thanks are due to Fundação para a Ciência e Tecnologia (FCT, Portugal) for funding the LAQV-REQUIMTE Research Unit (UIDB/50006/2020) and for financial support to project AMAZING (CIRCNA/BRB/0281/2019) and Dr. Cleópatra Caldera for yours opinions. The authors express their gratitude to Fundacao Oswaldo Cruz (FIOCRUZ), Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPq), Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior (CAPES), and Fundacao de Amparo a Pesquisa do Estado de Rondonia (FAPERO).

Ethical statement

International ethical guidelines for scientific papers were followed in the preparation of this manuscript.

References

[1] S.P. Mackessy, Introduccion to toxicology of columbrid snake venoms: bology, venoms and envenomation, J. Toxicol. Toxin Rev. 21 (2002) vii–xi.
[2] Y. Jia, B.A. Cantu, E.E. Sanchez, J.C. Perez, Complementary DNA sequencing and identification of mRNAs from the venomous gland of Agkistrodon piscivorus leucostoma, Toxicon 51 (8) (2008) 1457–1466.
[3] World Health Organization, Guide Lines for the production control and Regulation of Snake Antivenom Immunoglobulins, (20/03/2020), 2010 1–134.
[4] L. The, Snake-bite envenoming: a priority neglected tropical disease, Lancet 390 (10089) (2017) 2.
[5] J.M. Gutierrez, G. Solano, D. Pla, et al., Preclinical evaluation of the efficacy of Antivenoms for snakebite envenoming: state-of-the-art and challenges ahead, Toxins (Basel) 9 (5) (2017).
[6] A. Kasturiratne, A.R. Wickremasinghe, N. de Silva, et al., The global burden of snakebite: a literature analysis and modelling based on regional estimates of envenoming and deaths, PLoS Med. 5 (11) (2008), e218, .

[7] J.M. Gutierrez, Snake envenomation in Central America, in: S.P. Mackessy (Ed.), Handbook of Venoms and Toxins of Reptiles, CRC Press, Florida USA 2010, pp. 491–508.
[8] A. Ochoa-Avilés, O.S. Heredia-Andino, S. Escandón, et al., Viperidae Snakebites in Ecuador: A Review of Epidemiological and Ecological Aspects, Toxicon X 7, 2020.
[9] K.D. Zaqueo, A.M. Kayano, T.F. Domingos, et al., BbrzSP-32, the first serine protease isolated from Bothrops brazili venom: purification and characterization, Comp Biochem Physiol A Mol Integr Physiol 195 (2016) 15–25.
[10] S.L. Da Silva, C.A. Dias-Junior, P.A. Baldasso, et al., Vascular effects and electrolyte homeostasis of the natriuretic peptide isolated from Crotalus oreganus abyssus (North American Grand Canyon rattlesnake) venom, Peptides 36 (2) (2012) 206–212.
[11] A.M. Kayano, R. Simoes-Silva, P.S. Medeiros, et al., BbMP-1, a new metalloproteinase isolated from Bothrops brazili snake venom with in vitro antiplasmodial properties, Toxicon 106 (2015) 30–41.
[12] J.J. Calvete, Snake venomics: from the inventory of toxins to biology, Toxicon 75 (2013) 44–62.
[13] J.J. Calvete, P. Juarez, L. Sanz, Snake venomics, Strategy and applications, J Mass Spectrom 42 (11) (2007) 1405–1414.
[14] J.J. Calvete, D. Petras, F. Calderon-Celis, et al., Protein-species quantitative venomics: looking through a crystal ball, J Venom Anim Toxins Incl Trop Dis 23 (2017) 27.
[15] J.J. Calvete, L. Sanz, Y. Angulo, et al., Venoms, venomics, antivenomics, FEBS Lett. 583 (11) (2009) 1736–1743.
[16] R.M. Kini, Anticoagulant proteins from snake venoms: structure, function and mechanism, Biochem. J. 397 (3) (2006) 377–387.
[17] B. Lomonte, Lys49 myotoxins: emerging insights into their modes of action, Toxicon 177 (Suppl. 1) (2020) S5.
[18] B. Lomonte, Y. Angulo, L. Calderon, An overview of lysine-49 phospholipase A2 myotoxins from crotalid snake venoms and their structural determinants of myotoxic action, Toxicon 42 (8) (2003) 885–901.
[19] B. Lomonte, J. Fernandez, L. Sanz, et al., Venomous snakes of Costa Rica: biological and medical implications of their venom proteomic profiles analyzed through the strategy of snake venomics, J. Proteome 105 (2014) 323–339.
[20] B. Lomonte, J.M. Gutierrez, G. Borkow, et al., Activity of hemorrhagic metalloproteinase BaH-1 and myotoxin II from Bothrops asper snake venom on capillary endothelial cells in vitro, Toxicon 32 (4) (1994) 505–510.
[21] S.P. Mackessy, The field of reptile toxicology: sankes, lizards, and their venoms, in: S.P. Mackessy (Ed.), Handbook of Venoms and Toxins of Reptiles, CRC Press, Florida USA 2010, pp. 3–21.
[22] L. Sanz, H.L. GibbsS, S.P. Mackssy, J.J. Calvete, Venom proteomes of closely related Sistrurus rattlesnakes with divergent diets, J. Proteome Res. 5 (2006) 2098–2112.
[23] R.J.R. McCleary, T.S. Kang, R.M. Kini, Reptile venoms as platform for drug development, in: G.F. King (Ed.), Venoms to Drugs, Royal Society of Chemistry, Cambridge UK 2015, pp. 129–162.
[24] J.R. Almeida, A.L.V. Palacios, R.S.P. Patino, et al., Harnessing snake venom phospholipases A2 to novel approaches for overcoming antibiotic resistance, Drug Dev. Res. 80 (1) (2019) 68–85.
[25] L.A. Calderon, J.C. Sobrinho, K.D. Zaqueo, et al., Antitumoral activity of snake venom proteins: new trends in cancer therapy, BioMed Research International (2014) 203639 Article number.
[26] L.F.M. Izidoro, J.C. Sobrinho, M.M. Mendes, et al., Snake venom L-amino acid oxidases: Trends in pharmacology and biochemistry, BioMed Research International 2014 (2014) 196754 Article number.
[27] T.M. Abd El-Aziz, A.G. Soares, J.D. Stockang, Snake venoms in drug discovery: valuable therapeutic tools for life saving, Toxins 11 (564) (2019) 1–25.
[29] R. Simoes-Silva, J. Alfonso, A. Gomez, et al., Snake Venom, A Natural Library of New Potential Therapeutic Molecules: Challenges and Current Perspectives, Curr Pharm Biotechnol 19 (4) (2018) 308–335.
[30] J.R. Almeida, L.M. Resende, R.K. Watanabe, et al., Snake venom peptides and low mass proteins: molecular tools and therapeutic agents, Curr. Med. Chem. 24 (2017) 3254–3282.
[31] A.E. Dennis, J. Cao, Y. Hsu, et al., Phospholipase A2 enzymes: physical structure, biological function, disease implication, chemical inhibition, and therapeutic intervention, Chemical reviews 111 (2011) 6131–6185 (Lipid Biochemistry, Metabolism, and Signaling).
[32] S.L. Da Silva, E.G. Rowan, F. Albericio, et al., Animal toxins and their advantages in biotechnology and pharmacology, BioMed Research International 2014 (2014) 951561 Article ID.
[33] R. Simões-Silva, J. Alfonso, A. Gomez, et al., Snake venom, a natural library of new potential therapeutic molecules: challenges and current perspectives, Curr. Pharm. Biotechnol. 19 (2018) 308–335.
[34] T.S. Kang, D. Georgieva, N. Genov, et al., Enzymatic toxins from snake venom: structural characterization and mechanism of catalysis, FEBS J. 278 (23) (2011) 4544–4576.
[35] J.R. Almeida, M. Lancellotti, A.M. Soares, et al., CoaTx-II, a new dimeric Lys49 phospholipase A2 from *Crotalus oreganus abyssus* snake venom with bactericidal potential: insights into its structure and biological roles, Toxicon 120 (2016) 147–158.
[36] J.R. Almeida, B. Mendes, M. Lancellotti, et al., A novel synthetic peptide inspired on Lys49 phospholipase A2 from Crotalus oreganus abyssus snake venom active against multidrug-resistant clinical isolates, Eur. J. Med. Chem. 149 (2018) 248–256.
[37] J.R. Almeida, B. Mendes, R.S.P. Patino, et al., Assessing the stability of historical and desiccated snake venoms from a medically important Ecuadorian collection, Comp Biochem Physiol C Toxicol Pharmacol 230 (2020) 108702.
[38] J.R. Almeida, L.M. Resende, A.G. Silva, et al., Biochemical and functional studies of ColTx-I, a new myotoxic phospholipase A2 isolated from Crotalus oreganus lutosus (Great Basin rattlesnake) snake venom, Toxicon 117 (2016) 1–12.
[39] R.M. Kini, Excitement ahead: structure, function and mechanism of snake venom phospholipase A2 enzymes, Toxicon 42 (2003) 827–840.
[40] J.M. Gutierrez, L. Sanz, J. Escolano, et al., Snake venomics of the lesser Antillean pit vipers *Bothrops caribbaeus* and *Bothrops lanceolatus*: correlation with toxicological activities and immunoreactivity of a heterologous antivenom, J. Proteome Res. 7 (10) (2008) 4396–4408.
[41] Y. Jia, B. Ermolinsky, A. Garza, D. Provenzano, Phospholipase A2 in the venom of three cottonmouth snakes, Toxicon 135 (2017) 84–92.
[42] S. Marcussi, P.R. Santos, D.L. Menaldo, et al., Evaluation of the genotoxicity of Crotalus durissus terrificus snake venom and its isolated toxins on human lymphocytes, Mutat. Res. 724 (1–2) (2011) 59–63.
[43] D.C. Damico, T. Vassequi-Silva, F.D. Torres-Huaco, et al., LmrTX, a basic PLA (2) (D49) purified from Lachesis muta rhombeata snake venom with enzymatic-related antithrombotic and anticoagulant activity, Toxicon 60 (5) (2012) 773–781.
[44] D.C. Damico, M.A. da Cruz Hofling, M. Cintra, et al., Pharmacological study of edema and myonecrosis in mice induced by venom of the bushmaster snake (Lachesis muta muta) and its basic Asp49 phospholipase A(2) (LmTX-I), Protein J. 27 (6) (2008) 384–391.
[45] B.W. Dijkstra, J. Drenth, K.H. Kalk, Active site and catalytic mechanism of phospholipase A2, Nature 289 (5798) (1981) 604–606.
[46] B.W. Dijkstra, J. Drenth, K.H. Kalk, P.J. Vandermaelen, Three-dimensional structure and disulfide bond connections in bovine pancreatic phospholipase A2, J. Mol. Biol. 124 (1) (1978) 53–60.
[47] B.W. Dijkstra, K.H. Kalk, W.G. Hol, J. Drenth, Structure of bovine pancreatic phospholipase A2 at 1.7A resolution, J Mol Biol 147 (1) (1981) 97–123.
[48] H.M. Verheij, J.J. Volwerk, E.H. Jansen, et al., Methylation of histidine-48 in pancreatic phospholipase A2, Role of histidine and calcium ion in the catalytic mechanism, Biochemistry 19 (4) (1980) 743–750.
[49] R.A. Shipolini, G.S. Bailey, B.E. Banks, The separation of neurotoxin from the venom of Naja melanoleuca and the primary sequence determination, Eur. J. Biochem. 42 (1) (1974) 203–211.
[50] R.A. Shipolini, G.L. Callewaert, R.C. Cottrell, et al., Phospholipase a from bee venom, Eur. J. Biochem. 20 (4) (1971) 459–468.
[51] E.A. Dennis, Diversity of group types, regulation, and function of phospholipase A2, J. Biol. Chem. 269 (18) (1994) 13057–13060.
[52] C.H. Ernst, E.M. Ernst, Venomous Reptiles of the United States, Canada, and Northern Mexico Johns Hopkins University Press, Maryland USA 2011.
[53] H.K. Gloyd, R. Conant, Snakes of the Agkistrodon Complex: A Monographic Review, Society for the study of amphibians and reptiles, Ohio USA, 1990.
[54] J. Mitcheel, The Reptiles of Virginia Smithsonian, First Edition edition, Virginia USA, 1994.
[55] B. Lomonte, W.C. Tsai, J.M. Urena-Diaz, et al., Venomics of New World pit vipers: genus-wide comparisons of venom proteomes across Agkistrodon, J. Proteome 96 (2014) 103–116.
[56] L.M. Resende, J.R. Almeida, R. Schezaro-Ramos, et al., Exploring and understanding the functional role, and biochemical and structural characteristics of an acidic phospholipase A2, AplTx-I, purified from *Agkistrodon piscivorus leucostoma* snake venom, Toxicon 127 (2017) 22–36.
[57] J. Fernandez, J.M. Gutierrez, Y. Angulo, et al., Isolation of an acidic phospholipase A2 from the venom of the snake Bothrops asper of Costa Rica: biochemical and toxicological characterization, Biochimie 92 (3) (2010) 273–283.
[58] P.P. Marques, A. Esteves, M. Lancellotti, et al., Novel acidic phospholipase A2 from Porthidium hyoprora causes inflammation with mast cell rich infiltrate, Biochem Biophys Rep 1 (2015) 78–84.
[59] W. Martins, P.A. Baldasso, K.M. Honorio, et al., A novel phospholipase A2 (D49) from the venom of the Crotalus oreganus abyssus (North American Grand canyon rattlesnake), Biomed. Res. Int. 2014 (2014) 654170.
[60] Y. Jia, J. Villarreal, Phospholipase A2 purified from cottonmouth snake venoms display antibacterial effect against four representaive bacterial species, Toxicon 151 (2018) 1–4.
[61] E.A. Correa, A.M. Kayano, R. Diniz-Sousa, et al., Isolation, structural and functional characterization of a new Lys49 phospholipase A2 homologue from Bothrops neuwiedi urutu with bactericidal potential, Toxicon 115 (2016) 13–21.
[62] U.K. Laemmli, Cleavage of structural proteins during the assembly of the head of bacteriophage T4, Nature 227 (5259) (1970) 680–685.
[63] A.K. Calgarotto, D.C. Damico, L.A. Ponce-Soto, et al., Biological and biochemical characterization of new basic phospholipase A(2) BmTX-I isolated from Bothrops moojeni snake venom, Toxicon 51 (8) (2008) 1509–1519.
[64] A.A. de Moura, A.M. Kayano, G.A. Oliveira, et al., Purification and biochemical characterization of three myotoxins from Bothrops mattogrossensis snake venom with toxicity against Leishmania and tumor cells, Biomed. Res. Int. 2014 (2014) 195356.
[65] L. Romero, S. Marcussi, D.P. Marchi-Salvador, et al., Enzymatic and structural characterization of a basic phospholipase A(2) from the sea anemone Condylactis gigantea, Biochimie 92 (8) (2010) 1063–1071.
[66] M. Holzer, S.P. Mackessy, An aqueous endpoint assay of snake venom phospholipase A2, Toxicon 34 (10) (1996) 1149–1155.
[67] P. Edman, A method for the determination of amino acid sequence in peptides, Arch. Biochem. 22 (3) (1949) 475.
[68] H.M. Berman, J. Westbrook, Z. Feng, et al., The protein data bank, Nucleic Acids Res. 28 (1) (2000) 235–242.
[69] A. Waterhouse, M. Bertoni, S. Bienert, et al., SWISS-MODEL: homology modelling of protein structures and complexes, Nucleic Acids Res. 46 (W1) (2018) W296–W303.

[70] A. Bateman, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 47 (2019) D506–D515.
[71] Y.H. Pan, T.M. Epstein, M.K. Jain, B.J. Bahnson, Five coplanar anion binding sites on one face of phospholipase A2: relationship to Interface binding, Biochemistry 40 (2001) 609–617.
[72] D.L. Scott, S.P. White, Z. Otwinowski, et al., Interfacial catalysis: the mechanism of phospholipase A2, Science 250 (4987) (1990) 1541–1546.
[73] L.L.C. Schrodinger, The PyMOL Molecular Graphics System. Version 1.8, 2015.
[74] D.A. Case, Y. Ben-Shalom, S.R. Brozell, et al., Amber 18, Univ, California, San Francisco, 2018.
[75] J. Wang, R.M. Wolf, J.W. Caldwell, et al., Development and testing of a general amber force field, J. Comput. Chem. 25 (9) (2004) 1157–1174.
[76] C.I. Bayly, P. Cieplak, W.D. Cornell, P.A.A. Kollman, Well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model, J. Phys. Chem. 97 (1993) 10269–10280.
[77] C.C.J. Roothaan, New developments in molecular orbital theory, Rev. Mod. Phys. 23 (1951).
[78] J.A. Maier, C. Martinez, K. Kasavajhala, et al., ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB, J Chem Theory Comput 11 (8) (2015) 3696–3713.
[79] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, et al., Comparison of simple potential functions for simulating liquid water, J. Chem. Phys. 79 (1983) 926–935.
[80] J. Ryckaert, G. Ciccotti, H.J.C. Berendsen, Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes, J. Comput. Phys. 23 (1977) 327–341.
[81] M. Wiederstein, M.J. Sippl, ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins, Nucleic Acids Res. 35 (2007) 407–410.
[82] O.G. Berg, M.H. Gelb, M. Tsai, M.K. Jain, Interfacial enzymology: the secreted phospholipase A2-paradigm, Chem. Rev. 101 (2001).
[83] J.M. Gutierrez, B. Lomonte, Phospholipase A2 myotoxins from Bothrops snake venoms, Toxicon 33 (11) (1995) 1405–1424.
[84] I. Kudo, M. Murakami, Phospholipase A2, Tanpakushitsu Kakusan Koso 44 (8 Suppl) (1999) 1013–1024.
[85] I. Kudo, M. Murakami, Phospholipase A2 enzymes, Prostaglandins Other Lipid Mediat 68-69 (2002) 3–58.
[86] B. Mendes, J.R. Almeida, N. Vale, et al., Potential Use of 13-Mer Peptides Based on Phospholipase and Oligoarginine as Leishmanicidal Agents, Comparative Biochemistry and Physiology Part - C: Toxicology and Pharmacology 226, 2019.
[87] M. Terán, B. Lomonte, Actividad letal de seis venenos de serpientes de importancia médica en el Ecuador, Revista Ecuatoriana de Medicina y Ciencias Biologicas 37 (2016) 25–30.
[88] V.O. Zambelli, G. Picolo, C.A.H. Fernandes, et al., Secreted phospholipases A(2) from animal venoms in pain and analgesia, Toxins (Basel) 9 (12) (2017).
[89] D.C. Damico, L.G. Bueno, L. Rodrigues-Simioni, et al., Functional characterization of a basic D49 phospholipase A2 (LmTX-I) from the venom of the snake Lachesis muta muta (bushmaster), Toxicon 47 (7) (2006) 759–765.
[90] C.L. Guimaraes, L.S. Moreira-Dill, R.S. Fernandes, et al., Biodiversity as a source of bioactive compounds against snakebites, Curr. Med. Chem. 21 (25) (2014) 2952–2979.
[91] J. Fernandez, P. Caccin, G. Koster, et al., Muscle phospholipid hydrolysis by Bothrops asper Asp49 and Lys49 phospholipase A(2) myotoxins–distinct mechanisms of action, FEBS J. 280 (16) (2013) 3878–3886.
[92] K.C. Giannotti, E. Leiguez, V. Moreira, et al., A Lys49 phospholipase A2, isolated from Bothrops asper snake venom, induces lipid droplet formation in macrophages which depends on distinct signaling pathways and the C-terminal region, Biomed. Res. Int. 2013 (2013) 807982.
[93] C. UniProt, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 47 (D1) (2019) D506–D515.
[94] M. Murakami, I. Kudo, Phospholipase A2, J. Biochem. 131 (3) (2002) 285–292.
[95] J.A. Villar, F.T. Lima, C.L. Veber, et al., Synthesis and evaluation of nitrostyrene derivative compounds, new snake venom phospholipase A2 inhibitors, Toxicon 51 (8) (2008) 1467–1478.
[96] S.H. Andriao-Escarso, A.M. Soares, V.M. Rodrigues, et al., Myotoxic phospholipases A(2) in bothrops snake venoms: effect of chemical modifications on the enzymatic and pharmacological properties of bothropstoxins from Bothrops jararacussu, Biochimie 82 (8) (2000) 755–763.
[97] Y. Angulo, J.M. Gutierrez, A.M. Soares, et al., Myotoxic and cytolytic activities of dimeric Lys49 phospholipase A2 homologues are reduced, but not abolished, by a pH-induced dissociation, Toxicon 46 (3) (2005) 291–296.
[98] J.E. Burke, E.A. Dennis, Phospholipase A2 biochemistry, Cardiovasc. Drugs Ther. 23 (1) (2009) 49–59.
[99] J.E. Burke, E.A. Dennis, Phospholipase A2 structure/function, mechanism, and signaling, J. Lipid Res. 50 (Suppl) (2009) S237–S242.
[100] T.R. Costa, D.L. Menaldo, C.Z. Oliveira, et al., Myotoxic phospholipases A(2) isolated from Bothrops brazili snake venom and synthetic peptides derived from their C-terminal region: cytotoxic effect on microorganism and tumor cells, Peptides 29 (10) (2008) 1645–1656.
[101] A. de Roodt, J. Fernandez, D. Solano, B. Lomonte, A myotoxic Lys49 phospholipase A2-homologue is the major component of the venom of Bothrops cotiara from Misiones, Argentina, Toxicon 148 (2018) 143–148.
[102] J.M. Gutierrez, B. Lomonte, Phospholipases A2: unveiling the secrets of a functionally versatile group of snake venom toxins, Toxicon 62 (2013) 27–39.
[103] C. Araya, B. Lomonte, Antitumor effects of cationic synthetic peptides derived from Lys49 phospholipase A2 homologues of snake venoms, Cell Biol. Int. 31 (3) (2007) 263–268.
[104] R.B. Benati, T.R. Costa, M.D.C. Cacemiro, et al., Cytotoxic and pro-apoptotic action of MjTX-I, a phospholipase A2 isolated from Bothrops moojeni snake venom, towards leukemic cells, J Venom Anim Toxins Incl Trop Dis 24 (2018) 40.
[105] J.E. Burke, A.E. Dennis, Phospholipase A2 structure/function, mechanism, and signaling, Journal of Lipid Research April Supplement (2009) S237–S242.
[106] J.J. Hiu, M.K.K. Yap, Cytotoxicity of snake venom enzymatic toxins: phospholipase A2 and l-amino acid oxidase, Biochem. Soc. Trans. 48 (2) (2020) 719–731.
[107] B. Spolaore, J. Fernandez, B. Lomonte, et al., Enzymatic labelling of snake venom phospholipase A2 toxins, Toxicon 170 (2019) 99–107.
[108] C.N. Boeno, M.V. Paloschi, J.A. Lopes, et al., Inflammasome activation induced by a snake venom Lys49-phospholipase A2 homologue, Toxins (Basel) 12 (1) (2019).
[109] E.S. de Alvarenga, S.A. Silva, L.C. Barosa, et al., Synthesis and evaluation of sesquiterpene lactone inhibitors of phospholipase A2 from Bothrops jararacussu, Toxicon 57 (1) (2011) 100–108.
[110] J.P. Chippaux, Snake-bites: appraisal of the global situation, Bull. World Health Organ. 76 (5) (1998) 515–524.
[111] J. Laines, A. Segura, M. Villalta, et al., Toxicity of Bothrops sp snake venoms from Ecuador and preclinical assessment of the neutralizing efficacy of a polyspecific antivenom from Costa Rica, Toxicon 88 (2014) 34–37.
[112] L.H. Vélez-Alarcón, J.J. Real-Cott, K.J. Idrovo-Castro, et al., Caracterización del accidente por mordedura de serpiente antendidos en unidades de salud, zona 5, Ecuador, Revista Científica Digital INSPILIP 1 (2010) 1–16.
[113] V.S. Hanna, E.A.A. Hafez, Synopsis of arachidonic acid metabolism: a review, J. Adv. Res. 11 (2018) 23–32.
[114] J.A. Pereanez, L.M. Preciado, J. Fernandez, et al., Snake venomics, experimental toxic activities and clinical characteristics of human envenomation by Bothrocophias myersi (Serpentes: Viperidae) from Colombia, J. Proteome 220 (2020) 103758.
[115] K. Meirer, D. Steinhilber, E. Proschak, Inhibitors of the arachidonic acid cascade: interfering with multiple pathways, Basic Clin Pharmacol Toxicol 114 (1) (2014) 83–91.
[116] H. Yin, C.M. Havrilla, L. Gao, et al., Mechanisms for the formation of isoprostane endoperoxides from arachidonic acid. "Dioxetane" intermediate versus beta-fragmentation of peroxyl radicals, J Biol Chem 278 (19) (2003) 16720–16725.
[117] Y. Yamaguchi, Y. Shimohigashi, T. Chiwata, et al., Lys-49-phospholipases A2 as active enzyme for beta-arachidonoyl phospholipid bilayer membranes, Biochem. Mol. Biol. Int. 43 (1) (1997) 19–26.
[118] P. Davies, P.J. Bailey, M.M. Goldenberg, A.W. Ford-Hutchinson, The role of arachidonic acid oxygenation products in pain and inflammation, Annu. Rev. Immunol. 2 (1984) 335–357.
[119] J.C. Sobrinho, A.F. Francisco, R. Simoes-Silva, et al., Antimyotoxic activity of synthetic peptides derived from bothrops atrox snake gamma phospholipase A2 inhibitor selected by virtual screening, Curr. Top. Med. Chem. 19 (22) (2019) 1952–1961.
[120] J.C. Sobrinho, A.M. Kayano, R. Simoes-Silva, et al., Anti-platelet aggregation activity of two novel acidic Asp49-phospholipases A2 from *Bothrops brazili* snake venom, Int J Biol Macromol 107 (Pt A) (2018) 1014–1022.
[121] L.A. Somma, P. Fuller, A. Foster, *Agkistrodon piscivorus* (Lacepède, 1789): U.S. geological survey, nonindigenous aquatic species database, https://nas.er.usgs.gov/queries/FactSheet.aspx?SpeciesID=1197 2020. (Accessed 24 August 2020).
[122] J. Vindas, Y. Carrera, B. Lomonte, et al., A novel pentameric phospholipase A2 myotoxin (PophPLA2) from the venom of the pit viper Porthidium ophryomegas, Int. J. Biol. Macromol. 118 (2018) 1–8.
[123] S. Forst, J. Weiss, P. Blackburn, et al., Amino acid sequence of a basic Agkistrodon halys blomhoffii phospholipase A2. Possible role of NH2-terminal lysines in action on phospholipids of Escherichia coli, Biochemistry 25 (1986) 4309–4314.

5.3. Outcomes and Future Prospects

The undeniable duality surrounding snakes dates to ancient civilizations where they were both feared and praised. In this chapter, we have looked at snakes from two different perspectives: as extraordinary living libraries of bioactive compounds, and as the perpetrators of one of the leading causes of death in rural areas of Africa, Asia and Latin America [276].

A study on $sPLA_2$ focuses on its structure, chemistry, and on the docking of a synthetic substrate NOBA (4-nitro-3-octanoyloxybenzoic acid). By resolving the structure of $sPLA_2$ from *Agkistrodon piscivorus leucostoma* using homology modeling, it was possible to perform the docking of NOBA (4-nitro-3-octanoyloxybenzoic acid) on $sPLA_2$. NOBA is a synthetic substrate shown by experimental results that it is hydrolyzed by $sPLA_2$. The study also emphasizes the role of Ca^{2+} in the catalysis performed by this enzyme. Structural similarities were found between the enzyme's modeled structure and other snake $sPLA_2$ X-Ray structures available in the PDB database. Subsequent molecular dynamics (MD) simulations in explicit solvent allowed to construct a solvated, fully functional, pre-catalytic Michaelis Complex for the system. NOBA and active-site water molecules spontaneously adopted stable positions and established interactions in full agreement with the reaction mechanism proposed for the physiological substrate, suggesting that NOBA hydrolysis is an excellent model to study phospholipid hydrolysis. In this work, we increase knowledge surrounding the basic $sPLA_2$ (D49) from the *Agkistrodon piscivorus leucostoma* venom, in hopes of contributing to the study of the catalytic mechanism of $sPLA_2$. By characterizing the mechanism of such a key enzyme in the venom of vipers, we will be closer to finding an antidote for its nefarious effects.

CHAPTER 6. Science communication

As a general concept, science communication relies on conveying scientific information to an audience through a given platform. The way through which we communicate science will depend on many factors – the complexity of the subject matter, the type of audience, the platform itself, etc. – and therefore, it can be broadly classified as science "inreach" or science outreach.

Science "inreach" is what we, doctoral students, do from the early start of our careers. We communicate results to our colleagues at lab meetings, share our breakthroughs at conferences and poster sessions, and may even write literature reviews for a wider audience of fellow scientists and researchers. This type of communication is established between experts and implies a higher education level.

On the opposite side of the spectrum is science outreach. This serves the purpose of effectively communicating science-related news to the general public: research discoveries and their possible contribution to society. This should be done while respecting the ethics of science communication: honesty, precision, audience relevance, process transparency, and specification of uncertainty about conclusions [277].

Today's citizens learn about science through a myriad of channels and platforms. As doctoral students, we are at a vantage point for communicating science: we have enough expert knowledge to understand complex scientific processes and research, and we can translate the specifics of a research project into simple concepts, often with appealing images and text.

The following subchapters illustrate the science "inreach" and outreach carried out throughout the doctoral program.

6.1. Biocombustíveis, serão a solução?

Matilde F. Viegas[a], Maria J. Ramos[a], and Pedro A. Fernandes[a]

[a] LAQV@REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

Viegas, M. F., Ramos, M. J., Fernandes, P. (2018). ***Biocombustíveis, serão a solução?*** Revista de ciência elementar, 6(2):043. Casa Das Ciências. https://doi.org/10.24927/rce2018.043

REVISTA DE CIÊNCIA ELEMENTAR

Biocombustíveis, serão a solução?

Matilde Viegas,
Maria João Ramos,
Pedro Alexandrino Fernandes*

UCIBIO, REQUIMTE/Universidade do Porto
*pafernan@fc.up.pt

CITAÇÃO
Viegas, M., Ramos, J.M., Fernandes, P.A (2018) Biocombustíveis, serão a solução?, *Rev. Ciência Elem.*, V6(02):043. doi.org/10.24927/rce2018.043

EDITOR
José Ferreira Gomes, Universidade do Porto

EDITOR CONVIDADO
José Francisco Rodrigues, Universidade de Lisboa

RECEBIDO EM
27 de março de 2018

ACEITE EM
18 de maio de 2018

PUBLICADO EM
18 de junho de 2018

COPYRIGHT
© Casa das Ciências 2018. Este artigo é de acesso livre, distribuído sob licença Creative Commons com a designação CC-BY-NC-SA 4.0, que permite a utilização e a partilha para fins não comerciais, desde que citado o autor e a fonte original do artigo.

rce.casadasciencias.org

"A 21 de Fevereiro de 2018, o nível de poluição de partículas finas com diâmetro inferior a 10 μm, PM10, atingiu em Paris níveis alarmantes, com um pico correspondente ao dobro do limite máximo. A Organização Mundial de Saúde (OMS) recomenda não exceder este limite máximo por mais de três dias por ano. Com este recente pico, os parisienses já o ultrapassaram."

Em: *Un pic de pollution «grave» en Ile-de-France*, **Le Monde 22 de Fevereiro de 2018**

Esta notícia é um dos inúmeros exemplos dos efeitos do aquecimento global, termo com o qual estamos cada vez mais familiarizados, infelizmente. Este fenómeno é consequência de Gases de Efeito de Estufa, GEE (dióxido de carbono, metano e óxido de azoto, entre outros) que se acumulam na atmosfera, enclausurando o calor e tornando o planeta mais quente. A produção em larga escala de GEEs teve início na Revolução Industrial (1800)[1] como resultado da combustão de combustíveis fósseis (carvão, petróleo e gás natural) para produção de eletricidade, aquecimento e transporte.

A preocupação em torno do aquecimento global tornou-se um assunto constante na agenda política, sobretudo nas últimas duas décadas. Os primeiros esforços coletivos resultaram no Protocolo de Quioto, em 1997, cujo objetivo era a redução das emissões de GEEs assentando nos argumentos de que i) o aquecimento global era uma realidade; e ii) o CO_2 produzido por ações humanas era o seu grande causador[2]. O primeiro período do Protocolo terminou em 2012 e, por isso, em 2015, as Nações Unidas promoveram a assinatura do novo Acordo de Paris, o qual consiste numa atualização do antigo Protocolo de Quioto com o objetivo de "fortalecer a resposta global à ameaça das alterações climáticas, no contexto de desenvolvimento sustentável e com esforço para a erradicação da pobreza"[2]. O novo Acordo, ao contrário do de Quioto, estabelece metas tangíveis e quantitativas. O objetivo é manter o aumento da temperatura média global abaixo dos 2º Celsius face à temperatura global na era pré-industrial[3]. Para atingir tal meta, é necessário um esforço mundial, contudo, em 2017, com a saída dos EUA do Acordo de Paris, o segundo maior emissor de GEEs logo a seguir à China[4], perdeu-se uma enorme

REVISTA DE CIÊNCIA ELEMENTAR

contribuição. Além do peso evidente que as suas emissões têm para o estado atual do aquecimento global, em falta estarão também a diplomacia e o encorajamento proporcionados por esta força mundial.

No Acordo de Paris, uma das cláusulas exige a inclusão de energias renováveis (eólica, marítima, solar) e de biocombustíveis no mercado de energia, encorajando os países a atingir níveis mínimos da sua utilização. As metas estabelecidas para cada país vão ao encontro ao seu mercado e às matérias-primas disponíveis em cada território, favorecendo o desenvolvimento da economia.

De uma forma generalista, os biocombustíveis englobam combustíveis produzidos através de processos biológicos, como a agricultura e digestão anaeróbica, em vez de produzidos a partir de processos geológicos, como os necessários para a formação de combustíveis fósseis. Isto inclui o biodiesel (produzido a partir de óleos e gorduras orgânicas) e o bioetanol (álcool produzido através de fermentação microbiana de açúcares, seguido de destilação e desidratação), os quais serão o maior foco deste artigo. Duas grandes vantagens são apresentadas como motivadoras para o uso de biocombustíveis: fontes ilimitadas de matérias-primas renováveis, ao contrário dos combustíveis fósseis, dependentes das reservas geológicas existentes; além disso, para a produção destes biocombustíveis, a matéria-prima provém, na sua maioria, da agricultura e desperdício tornando a sua pegada de carbono neutra (o dióxido de carbono libertado durante a sua combustão é igual à quantidade assimilada durante a fotossíntese)[5].

Embora o uso do biodiesel e bioetanol seja, de uma forma abrangente, positivo, é necessário reconhecer as falhas de alguns processos e a oportunidade que lobistas e sectores económicos poderão encontrar nos biocombustíveis. É por isso que o Acordo de Paris e outros conjuntos de políticas são essenciais para a implementação e regulamentação de energias renováveis e biocombustíveis. Ao lado do desenvolvimento científico e investimento monetário, é necessária uma estrita política mundial que estabeleça metas de consumo dos biocombustíveis e energias renováveis no mercado, além de um crucial controlo da indústria.

A partir da década de 70, foram feitos os primeiros investimentos nos chamados biocombustíveis de primeira geração (1G). Estes fazem uso de colheitas de produtos agrícolas, como a soja, cana-de-açúcar, milho, óleo de palma, entre outros, para a produção de biodiesel e bioetanol através de processos de trans-esterificação ou fermentação. A sua disponibilidade e facilidade de processamento permitiram a sua inserção e mistura nos combustíveis atualmente vendidos ao público[6]. O biodiesel pode ser diretamente utilizado nos veículos, sem qualquer alteração, no entanto, a sua utilização pode danificar o motor dos automóveis. Quanto ao bioetanol, este está presente na gasolina disponível em qualquer gasolineira, em misturas contendo até 7,5% de bioetanol no volume total.

Os biocombustíveis 1G fazem uso da colheita agrícola para fins energéticos, o que gera opiniões desfavoráveis por parte de ambientalistas (pelo uso exaustivo da terra agrícola sem compensação, desflorestação e perda de biodiversidade) e problemas éticos, tais como os levantados por responsáveis da UN, como o Dr. Jean Ziegler, que atribui aos biocombustíveis a culpa pela contínua inflação dos preços de alimentos, exacerbando a atual crise de fome mundial- “São necessários 352Kg de milho para encher um depósito de 50L de um carro. Uma criança na Zâmbia ou no México, onde o

REVISTA DE CIÊNCIA ELEMENTAR

milho é um elemento básico da dieta, poderia viver com esta quantidade de milho um ano inteiro"[7]. Atualmente, os maiores produtores destes biocombustíveis de primeira geração são os EUA e o Brasil, utilizando cerca de parte da produção global de cereais e óleos vegetais para a produção de bioetanol e biodiesel, causando um ambiente de profunda preocupação em volta da sustentabilidade de produção de energia a médio/longo prazo. De forma a controlar a inflação no preço dos alimentos, a EU estabeleceu um limite de 7% para o uso de biocombustíveis de primeira geração como combustível de transportação[8].

A segunda geração de biocombustíveis (2G) usa biomassa, isto é, qualquer fonte de carbono orgânico que é rapidamente renovado no ciclo de carbono. Os biocombustíveis de segunda geração almejam um uso ponderado de desperdícios da indústria madeireira e agrícola, ou de restos de colheitas agrícolas assim que estas preencham o seu propósito alimentar, evitando interferência com o mercado alimentar e não influenciando os preços dos alimentos básicos[5].

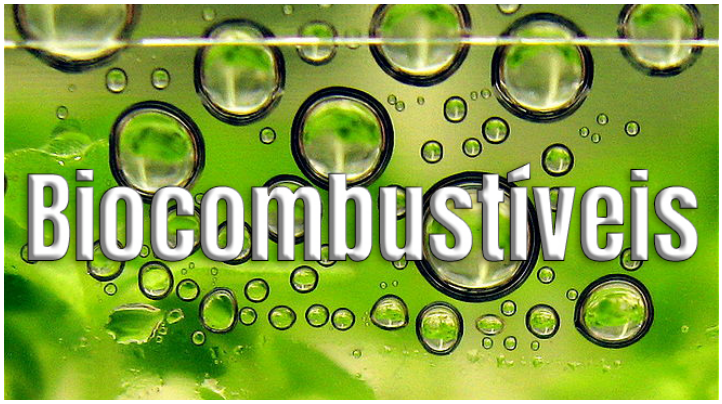


FIGURA 1. Biocombustíveis.

A produção de arroz atingiu o seu pico em 2017, gerando 1140 milhões de toneladas de espigas de arroz, cujo destino é a incineração, compostagem, ou é simplesmente depositada em aterros. Contudo, com a 2G de biocombustíveis, é encontrado potencial energético neste desperdício agrícola[9]. O seu potencial deve-se à matéria lignocelulósica – celulose, hemicelulose e lignina – presente em todas as plantas, incluindo as espigas, em diferentes percentagens. Este material composto é essencial para a estrutura e resistência das plantas: a hemicelulose é a matriz que rodeia o esqueleto de celulose, enquanto a lignina funciona como uma camada protetora. As possibilidades associadas a este tipo de desperdício são inegáveis: a sua abundância e facilidade de obtenção tornam-no muito apelativo. Contudo, o gasto na recolha e transporte desta matéria suplanta os custos de produção e o seu uso continuado poderá contribuir para a erosão do solo. Além disso, para o seu aproveitamento para produção de biocombustíveis, é necessário um pré-tratamento para hidrolisar as ligações covalentes entre os três constituintes da matéria lignocelulósica e

REVISTA DE CIÊNCIA ELEMENTAR

permitir a posterior fermentação anaeróbia da celulose, produzindo bioetanol. Este pré-tratamento é visto como outra das desvantagens do processo e está ainda a ser estudado e otimizado, pois de momento as alternativas disponíveis encarecem o produto final e, em alguns dos casos, dependem do uso de químicos poluentes. A esperança está então depositada no estudo de enzimas hidrolíticas, encontradas em organismos termófilos, que serão capazes de aguentar as difíceis condições de tratamento. Contudo, a dificuldade em escalar o processo de fermentação é uma das desvantagens e dos fatores limitantes para a 2G de biocombustíveis, em grande parte devido à dificuldade de monitorizar as condições de fermentação (oxigénio, pH, nutrientes, etc)[9].

A terceira geração (3G) de biocombustíveis foi reconhecida como a melhor alternativa em comparação com primeira e segunda gerações, pois, surpreendentemente, faz uso de algas[5]. A sua enorme disponibilidade por todo o planeta e o facto de não interferirem com os mercados alimentares são os grandes embaixadores desta matéria-prima como o futuro dos biocombustíveis. Ainda que não seja uma matéria infinita, é a que de momento mais se assemelha, por existirem 70 000 espécies (30 000 documentadas)[10]. Estudos conduzidos atualmente estão a explorar a viabilidade energética de 30 espécies de algas, explorando a sustentabilidade, facilidade e escalabilidade da sua produção. Outras vantagens incluem a baixa manutenção necessária para o seu crescimento (luz, dióxido de carbono e nutrientes como azoto, fósforo e potássio); a facilidade com que produzem lípidos e carboidratos que podem posteriormente ser processados em diferentes biocombustíveis e outros coprodutos; a frequência com que podem ser colhidas ao longo do ano (ao contrário de outras colheitas convencionais); a baixa percentagem de hemicelulose, e a praticamente ausente lignina, facilitam radicalmente o seu pré-processamento pois a celulose é mais facilmente extraída. As desvantagens desta matéria incluem o enorme investimento inicial para a construção de infraestruturas para o seu cultivo, que pode ser ou ao ar livre (em tanques de grande dimensão) ou em foto-bioreatores (sistemas fechados altamente controlados)[11]. Prevê-se que em 2030, o preço de produção de combustível a partir de algas[12] será 0,65 €/litro de combustível, valor muito próximo do atual preço de produção do litro de gasolina. Comparativamente aos custos de produção de outros biocombustíveis, o mais barato continua a ser o bioetanol produzido a partir de milho (1G), 0,17 €/litro, seguido de combustíveis celulósicos (2G) a 0,32 €/litro[6].

É evidente que serão necessários esforços políticos e científicos, investimento e regulamentação adequados, para encontrar o caminho ideal para travar o aquecimento global. Em todas as gerações de biocombustíveis apresentadas, existem fatores que afunilam as potencialidades destas alternativas, indicando que a solução passa por criar um consórcio entre as várias possibilidades de forma a tirar o máximo partido sem danificar o planeta ou piorar a crise da fome. O próximo passo passa por substituir os biocombustíveis de 1G por soluções com maior escalabilidade. Acreditamos que nas próximas décadas serão cruciais, e depositamos esperança na investigação científica.

REVISTA DE CIÊNCIA ELEMENTAR

REFERÊNCIAS

[1] Changes since the Industrial Revolution - American Chemical Society, acesso em 16 de abril de 2018.

[2] United Nations Treaty Collection.

[3] LEWIS, S. L., The Paris Agreement Has Solved a Troubling Problem. *Nature*, 532 (7599), 283–283, 2016.

[4] Hai-bin, Z. *et al.*, W. Science Direct U.S. Withdrawal from the Paris Agreement : Reasons , Impacts , and China ' s Response. *Adv. Clim. Chang. Res.*, 8 (4), 220–225, 2017.

[5] Aro, E., From First Generation Biofuels to Advanced Solar Biofuels. *Ambio*, 45 (1), 24–31, 2016.

[6] CLARK, J. H. *et al.*, Green Chemistry, Biofuels, and Biorefinery. *Annu. Rev. Chem. Biomol. Eng.*, 3 (1), 183–207, 2012.

[7] Ziegler, J., Biofuels a big cause of famine, Globalization, *DW.*, acesso em 27 de março de 2018.

[8] EU agrees plan to cap use of food-based biofuels, acesso em 16 de abril de 2018.

[9] SATLEWAL, A. *et al.*, Rice Straw as a Feedstock for Biofuels: Availability, Recalcitrance, and Chemical Properties. *Biofuels, Bioprod. Biorefining*, 12 (1), 83–107, 2018.

[10] GUIRY, M. D., How many species of algae are there?, *J. Phycol.*, 48 (5), 1057–1063, 2012.

[11] DALGLEISH, P., Sustainability Research Project: Research of Suitable Locations, Design and Operation of Microalgae Production Plants for Biofuel's. *Nat. Resour.*, 08 (11), 671–708, 2017.

[12] The DOE's shifting worldview for biofuels deployment, now through 2030 : Biofuels Digest, acesso em 16 de abril de 2018.

View publication stats

6.2. Proteínas Virais no *Protein Data Bank*

Pinto, A. V., Oliveira, A. N.L., Jerves, C., Mendonça, F. L., Coutinho, I. F. M., Coimbra, J.T.S., Sousa, J. P.M., Cabrera-Trujillo, J. J., Teixeira, L., Viegas, M. F., Ferreira, P., Paiva, P., Fonseca, R., Neves, R. P.P., Magalhães, A. L., Fernandes, P. A., Ramos M. J.

LAQV@REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

---

Pinto, A. V., Oliveira, A. N.L., Jerves, C., Mendonça, F. L., Coutinho, I. F. M., Coimbra, J.T.S., Sousa, J. P.M., Cabrera-Trujillo, J. J., Teixeira, L., **Viegas, M. F**., Ferreira, P., Paiva, P., Fonseca, R., Neves, R. P.P., Magalhães, A. L., Fernandes, P. A., Ramos M. J. (2020). ***Proteínas Virais no* Protein Data Bank.** Revista de ciência elementar, 8(2):022. Casa Das Ciências. https://doi.org/10.24927/rce2018.043

REVISTA DE CIÊNCIA ELEMENTAR

Proteínas Virais no *Protein Data Bank*

Alexandre V. Pinto, Ana N.L. Oliveira, Carola Jerves, Filipa Lopes de Mendonça, Inês F.M. Coutinho, João T.S. Coimbra, João P.M. Sousa, Jorge J. Cabrera-Trujillo, Luís Teixeira, Matilde F. Viegas, Pedro Ferreira, Pedro Paiva, Rita Fonseca, Rui P.P. Neves, Alexandre L. Magalhães, Pedro A. Fernandes, Maria João Ramos
DQB/LAQV/Universidade do Porto

CITAÇÃO
Pinto, A. *et al.* (2020) Proteínas Virais no *Protein Data Bank*, ***Rev. Ciência Elem.***, V8(02):022. doi.org/10.24927/rce2020.022

EDITOR
José Ferreira Gomes, Universidade do Porto

EDITOR CONVIDADO
Pedro A. Fernandes, Universidade do Porto

RECEBIDO EM
19 de maio de 2020

ACEITE EM
20 de maio de 2020

PUBLICADO EM
30 de junho de 2020

COPYRIGHT
© Casa das Ciências 2020. Este artigo é de acesso livre, distribuído sob licença Creative Commons com a designação CC-BY-NC-SA 4.0, que permite a utilização e a partilha para fins não comerciais, desde que citado o autor e a fonte original do artigo.

rce.casadasciencias.org

Descoberto um novo vírus, o SARS-CoV-2, em quatro meses foi já possível determinar 182 estruturas de um número muito reduzido de proteínas deste vírus, isoladas ou ligadas a outras entidades, com diferentes técnicas e resoluções. Apresentam-se visualizações adaptadas de algumas proteínas com estruturas já depositadas no Banco de Dados de Proteínas (*Protein Data Bank*) por cientistas de todo o mundo e através do qual todos colaboram.

Estruturas 3D de proteínas

As proteínas são macromoléculas constituídas fundamentalmente por aminoácidos, que fazem igualmente parte de todos os sistemas biológicos como sejam os vírus, as bactérias, as leveduras, as plantas, as moscas e os animais. O nome "proteína" vem da palavra grega *proteos* que significa "primário" ou "primeiro lugar". Nas proteínas são reconhecidos 4 níveis de estruturas esquematizados na FIGURA 1.

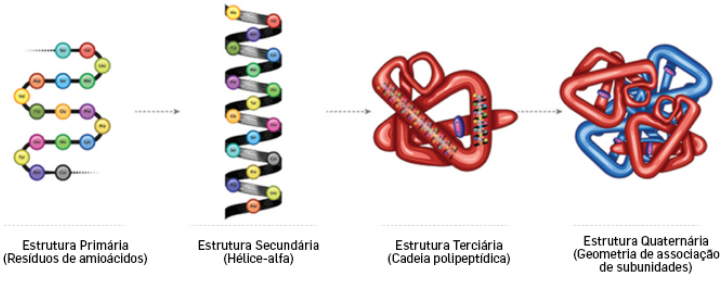


FIGURA 1. Os quatro níveis de estruturas das proteínas.

O nível mais simples da estrutura proteica, a estrutura primária, é simplesmente a sequência de aminoácidos de uma cadeia de polipéptidos. Segue-se a estrutura secundária que se refere a motivos repetitivos dentro da proteína tais como hélices ou folhas, que se formam devido a pontes de hidrogénio formadas entre os átomos da cadeia principal da proteína. A estrutura terciária (muitas vezes referida como tridimensional ou 3D) reflete o

REVISTA DE CIÊNCIA ELEMENTAR

enrolamento da cadeia de aminoácidos ou, muito simplesmente, a sua conformação global. Finalmente, a estrutura quaternária das proteínas refere-se ao modo como as várias cadeias proteicas (quando existentes) se dispõem entre si.

O Banco de Dados de Proteínas – *Protein Data Bank* (PDB)

O *Protein Data Bank* (PDB)[1] é um repositório de informação estrutural 3D de macromoléculas biológicas, nas quais se inserem as proteínas. Estas moléculas para além de fazerem parte da vida de todos os organismos como sejam as bactérias, plantas e animais, constituem também uma parte fundamental dos vírus.

O PDB foi estabelecido em 1971 no *Brookhaven National Laboratory*, EUA, contendo, originalmente, 7 estruturas. Neste momento, contém 164 174 estruturas macromoleculares biológicas e este número aumenta diariamente. O *Research Collaboratory for Structural Bioinformatics* (RCSB) tornou-se responsável pela gestão do PDB em 1998. Posteriormente, em 2003, foi criada a wwPDB ou seja a organização que supervisiona atualmente o PDB, tendo por função a manutenção de um único arquivo de dados estruturais macromoleculares do PDB, livre e publicamente disponível para a comunidade global. A informação estrutural que figura no PDB é normalmente obtida por biólogos e bioquímicos de todo o mundo, maioritariamente através de experiências de cristalografia de raios X, espetroscopia de ressonância magnética nuclear (RMN), crio-microscopia eletrónica (crio-EM), difração de neutrões ou mesmo modelação molecular. Os autores destas experiências depositam os seus resultados no arquivo do PDB sendo o seu conteúdo acessível livremente na internet através do site *www.rcsb.org*. As estruturas que figuram no PDB vão desde proteínas e ácidos nucleicos até máquinas moleculares complexas como o ribossoma. É importante compreender que a função de uma proteína está diretamente ligada à sua estrutura 3D e que o seu estudo tem um papel fundamental na saúde humana, na doença e na descoberta de fármacos.

As proteínas do vírus da COVID-19 no PDB

SARS-CoV-2 é o vírus causador da pandemia COVID-19[3]. Os vírus, em geral, e este, em particular, são estruturas simples quando comparados com todos os organismos vivos. Assim, o vírus SARS-CoV-2 é constituído por uma membrana viral, uma cadeia simples de RNA e quatro proteínas estruturais denominadas proteína E (do inglês *Envelope*), proteína M (do inglês *Membrane*), proteína N (do inglês *Nucleocapside*) e proteína S (do inglês *Spike*), para além de várias outras proteínas codificadas no RNA. A proteína N protege o material genético e as três proteínas, S, M e E, estruturam o envelope viral[2]. Apesar de só muito recentemente se terem iniciado estudos de estruturas 3D das proteínas do SARS-CoV-2, o PDB contém já 182 entradas obtidas nos últimos 4 meses que, de uma forma ou de outra, se relacionam com aquele número muito reduzido de proteínas, isoladas ou ligadas a outras entidades, com diferentes técnicas e resoluções.

Arte na Ciência com tema nas proteínas do vírus da COVID-19

A ciência e a arte são frequentemente consideradas como disciplinas distintas, sendo que, de um modo geral, se considera que uma pessoa não pode interessar-se seriamente por ambas a não ser que o interesse por uma esteja relacionado, de alguma forma, com o trabalho na outra. No entanto, na realidade, os cientistas participam e produzem arte a todos os níveis e em todos os meios. De acordo, com esta visão do mundo, já Albert Einstein afirmou

REVISTA DE CIÊNCIA ELEMENTAR

que "*os grandes cientistas são também grandes artistas*". Assim, e de acordo com as palavras deste grande mestre, tentamos encontrar dentro de nós os artistas que poderiam relacionar as proteínas do vírus que provoca a tão indesejada COVID-19, com algo harmonioso e visualmente agradável que, simultaneamente, dê a conhecer um pouco da maquinaria microscópica que transtorna a vida. Com as imagens que se seguem, conseguimos vislumbrar pequenos detalhes, pequenas imagens, pequenas composições desta sinfonia e, através delas, imaginar a inimaginável simplicidade sobre a qual o SARS-CoV-2 está construído.

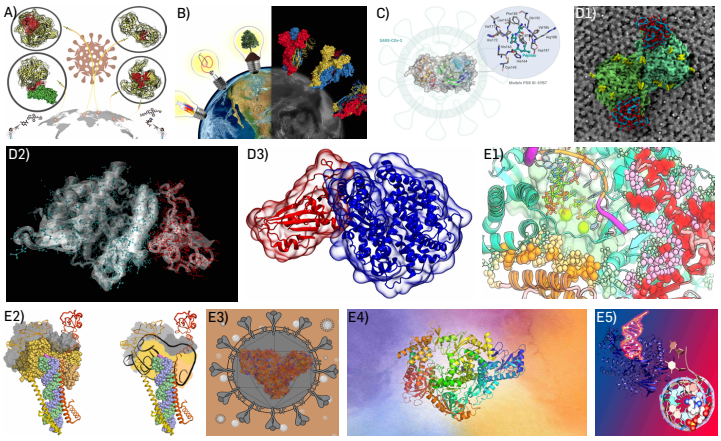


FIGURA 2. A) Coronavírus SARS-CoV-2. B) Proteína S do SARS-CoV-2. C) Protease do SARS-CoV-2. D1) a D3) Três visões do complexo Proteína S - hACE2. E1) a E5) Cinco visões da RNA polimerase do SARS-CoV-2, responsável pela replicação do seu genoma, essencial para a sua propagação. Estas imagens fazem parte da exposição virtual "Moléculas Magníficas - O Mundo Contra a COVID-19", organizado pelo Laboratório de Bioquímica Computacional da FCUP e pela Casa das Ciências. Esta exposição pode ser visitada em https://www.casadasciencias.org/artigo/o-mundo-contra-o-sars-cov-2.

As coordenadas 3D retiradas do PDB das estruturas proteicas deste vírus fornecem-nos apenas as geometrias que posteriormente trabalhamos para criarmos visualizações não só apelativas como também informativas das funções que as estruturas das proteínas comandam. Na realidade, as estruturas das proteínas estabelecem as bases das suas interações com outras moléculas do organismo e, por conseguinte, determinam as suas funções. Baseadas neste facto tão importante, as imagens aqui apresentadas tentam dar a conhecer ao leitor a maquinaria atómica do vírus através das suas proteínas, informar sobre as suas funções e transmitir uma consciência remota da dimensão do que sobre elas ignoramos e que, provavelmente, nunca chegaremos a compreender completamente.

REVISTA DE CIÊNCIA ELEMENTAR

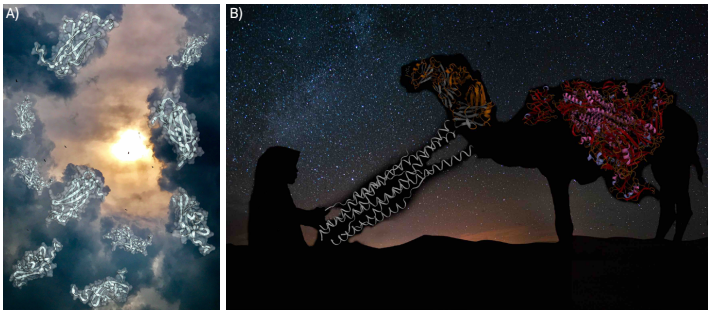


FIGURA 3. A) Proteína 9B da superfamília das imunoglobinas – IgSF9b. B) Proteína S do MERS (Síndrome Respiratório do Médio Oriente).

No entanto, e para além das suas funções artísticas já aqui documentadas, estas visualizações do mundo microscópico podem ser essenciais para a investigação de quem trabalha com estes sistemas biológicos. Como exemplo, examine-se a imagem C da FIGURA 2 deste texto. Ela representa a protease principal do vírus SARS-CoV-2 (MPro), responsável pela clivagem das poliproteínas do mesmo. Durante este processo, são libertadas proteínas virais funcionais e enzimas-chave para a replicação do vírus. Assim, o bloqueio da MPro impedirá a continuidade do SARS-CoV-2. E, nesse sentido, esta mesma imagem está a ser utilizada para estabelecer o mecanismo catalítico da MPro, tendo como base o centro ativo da proteína que se encontra em destaque à direita da imagem e que engloba todos os resíduos de aminoácidos importantes a considerar. Outras imagens têm uma finalidade pedagógica, embora não investigativa, ou ainda política, como seja a imagem A da FIGURA 2 que tenta chamar a atenção para as representações 3D de quatro proteínas, alvos promissores de fármacos contra o vírus SARS-CoV-2 causador da COVID-19. Na fila de cima, à esquerda, está a RNA polimerase dependente de RNA, que é indispensável para a replicação do vírus *in vivo*; na fila de cima, à direita, temos a protease principal (MPro) do SARS-CoV-2, já considerada no exemplo anterior; na fila de baixo, à esquerda, está representada a glicoproteína *spike* (S, espícula) fundamental para a entrada do vírus nas células do hospedeiro; e representada a verde temos a enzima conversora da angiotensina 2 (hACE2) que é o recetor à qual a glicoproteína S se liga nas células humanas; por último, na fila de baixo, à direita, está representada a protease do tipo papaína (PLpro) que, tal como a MPro, tem como função processar poliproteínas em proteínas funcionais, ajudando também o vírus a evadir-se da resposta imune do hospedeiro. Ainda na mesma imagem vemos também uma cientista, que após semanas de estudo científico, dispara moléculas com potencial terapêutico com uma arma cronicamente subfinanciada.

REFERÊNCIAS

[1] BERMAN, H.M. *et al.*, The Protein Data Bank, Nucleic Acids Research, 28, 235-242, 2000.

[2] LI, G. *et al.*, *Coronavirus infections and immune responses*, J. Med. Virology, 92, 424-432, 2020.

[3] ANDERSEN, K.G. *et al.*, *The proximal origin of SARS-CoV-2*, Nature Medicine, 26, 450–452, 2020.

View publication stats

6.3. Protocol for QM/MM Workshop at 2019 Dynapeutics Summer School

Sousa, J. P.M., Viegas, M. F., Fernandes, P. A., Ramos M. J.

LAQV@REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

DYNAPEUTICS INTERNATIONAL SUMMER SCHOOL 2019

The Reaction Mechanism of the DszB Enzyme - Protocol for the Practical Course on QM/MM Calculations- 3rd October 2019

Matilde Viegas · João Sousa · Pedro Alexandrino Fernandes · Maria João Ramos

Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Portugal

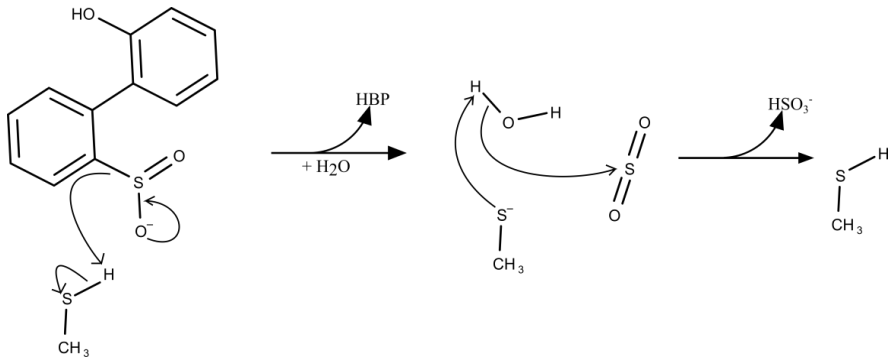
INTRODUCTION:

In this course the students will learn the basic procedures to determine a reaction mechanism of an enzyme. Enzymes constitute ideal systems to apply QM/MM methods because they have a small region (the active site and substrate/cofactors) where electronic rearrangements (in this case chemical reactions) take place, and a very large enzymatic molecular scaffold, with thousands of atoms, plus the solvent, all making electrostatic interactions with the active site but having no significant electronic rearrangements. In such scenario, the active site, substrate and cofactors may be described with quantum mechanics and the remaining system with classical mechanics.

DszB is a very interesting enzyme, with 365 residues, that participates in the sulfur metabolism of several bacteria and that has a great biotechnological potential due to its ability to remove sulfur from organic compounds at room temperature and pressure. In order to use this enzyme in industrial applications its catalytic mechanism has to be known, so that mutations can be rationally predicted and introduced to tune its activity and reactivity to the desired levels.

The catalytic mechanism is proposed to follow two possible pathways: either a nucleophilic addition or an electrophilic substitution pathway. Here we will investigate the feasibility of the later. For that we will calculate the activation free energy of the rate-determining step and see if it matches the experimental value. If this happens the simulated mechanism is probably the correct one.

According to this mechanism, the reaction starts with the protonation of a C carbon of the substrate, with elimination of SO_2, and in a second step SO_2 is attacked by a water molecule giving origin to HSO_3^- (observed in solution) and restoring the original active site protonation state (**Scheme 1**).



Scheme 1- Electrophilic aromatic substitution mechanism proposal for DszB reaction mechanism.

THE PROTOCOL:

The study consists in the division of the enzyme into a Quantum Mechanics and a Molecular Mechanics layer, a geometry optimization of the initial reactant state, previously modeled from the pdb structure 2DE3, and the determination of an approximate transition state structure for the first reaction step of the mechanism.

In the folder **Files**, you will find all the necessary files for this workshop.

I. Defining the QM and MM layers

Open a terminal window and launch the molecular visualization software VMD.

Load the initial reactant structure:

- vmd **Reactant.pdb**

After loading the reactant structure, identify the substrate and the residues that are most important for the first, rate-determining step of this reaction. The name of the substrate in the PDB file is OBP and the important residue are Ser6, Cys8, His41, Arg51 and Gly54.

Question: why these residues are expected to be the most important for the reaction?

To prepare our system for a QM/MM calculation, we start by defining which atoms will be described by quantum mechanics (the QM layer) and by classical mechanics (the MM layer). For that we run the software VMD molUp and load the **QMMM_prep.com** file. (Note: *.com files are the input files for the software Gaussian16. These files contain information on the molecule structure and instructions that define the calculation to be done). In this file we have worked the x-ray structure, in order to add the missing hydrogen atoms and add a layer of solvent.

- vmd > Extensions > MolUP

On the tab *"Model"*, select the atoms for the QM layer. To do so, you can go to *Model* > *Layer* tab, use the molUP selection tools *(Atom selection (Change ONIOM layer)"*) and pick the atoms by typing the selection "resid 6 8 51 54 or resname OBP" for the QM layer (initially, all atoms are defined as Low Layer/ MM atoms).

Challenge: Based on what you have learned in the morning lectures, try to choose the appropriate QM layer.

Then go to the *"Freeze"* tab and select all the protein and substrate atoms to "unfreeze" them (frozen atoms are defined by number *"-1"* and unfrozen atoms are defined by number *"0"*), to do so type "protein or resname OBP" on the atom selection box, select number 0 and *"Apply"*.

Check visually the QM and MM layers on the molecular representations (check *"High layer"* box at the bottom of the molUP interface).

On the tab "*Input"* introduce the theoretical level to be employed in the QM layer. B3LYP is a good choice for geometry optimizations and the basis set 6-31G(d) offers a good compromise between the accuracy of the QM layer geometry and the time needed for the calculation. The MM layer will be treated with AMBER force field. To insert these specifications in the input file, go to *"Calculations"* box and type:

oniom(B3LYP/6-31G(d):AMBER=softfirst) geom=connectivity opt

To save the input file go to *"File"* and select *"Save",* save the file as **QMMM.com**.

II. Optimize the geometry of the Reactant.

Launch the QM/MM calculation on the terminal window:

➢ g16 **QMMM.com** &

This calculation can take many hours, or days. To avoid spending too much time waiting, you can stop the calculation and copy the file **initial_opt.log** from the folder **outputs** to the folder you are working in. This file contains the result of the calculation you have just stopped.

On the molUP interface, load the file **initial_opt.log**. Load all structures to see each step of the geometry optimization. Check the box *"High Layer"* and uncheck the box *"All"* on the molUP interface. Press "*Play* " on the VMD Main Window and visually inspect how the structure relaxes throughout the optimization.

> *Question: The most important interactions for the chemical reaction are preserved through the geometry optimization?*

Save the last structure of the optimization calculation (i.e. the optimized structure), as an input file for Gaussian16, with the name of **PEP.com**, so we can proceed with the study of the first step of the reaction mechanism.

III. Calculate a Potential Energy Profile and Identify an Approximate Transition State Structure.

To study a reaction step, you need to generate a good guess of the transition state structure. The usual method is to calculate a Potential Energy Surface (PES). A PES is a function that has as arguments some, or all, of the atomic coordinates of the system, and as a value the energy of the system. The specific coordinates of each PES are chosen depending on the studied phenomenon. For a chemical reaction, the PES usually involves the distances between reacting atoms. If a PES has as argument a single coordinate, then it is usually called a Potential Energy Profile (PEP).

In the case we are studying, it is reasonable to assume that the distance between the SO_2 bound C atom and the electrophilic proton from Cys27 thiol group may be a reasonable representation of the reaction coordinate. As such, we will calculate a PEP along this distance and see if the PEP increases up to a maximum in energy and comes down again, defining a transition state structure and a product along the PEP. The energy maximum should be close to the true transition state structure along the true reaction coordinate. For that we will constrain the SH—C distance to a set of specific points and relax the remaining system, to calculate the energy along the SH—C distance. Such calculation is usually named as a *linear scan*.

On the molUP window, go to the "*Input"* section. Change the calculation keywords to perform the linear scan. In this case, it is important to introduce the "ModRed" option associated with the "opt" keyword in order to instruct Gaussian16 to perform a linear scan:

Click on "*Other information (Parameters, ModRedundant…)"* menu and select "*ModRedundant Editor"*.

On the pop-up window, click on "*Pick"* to select the pair of atoms whose distance you want to scan.

After you need to define the points along the SH—C distance that will be calculated. You should start to measure the initial and "expected" final SH—C distances, to see the distance range you need to span. Afterwards you need to assume a reasonable step size (the distance between points), not too fine so that the calculation does not takes too long, and not too large, so that the PEP has good definition. Generally, for a first linear scan exploration, points separated by 0.1 Å is a good choice.

Introduce the number of points that will be used in the PES. Provide the size of the step in Å. Positive numbers will increase the distance between both atoms whereas negative numbers will short their distance. Click in "*Apply"*. On the "*Other information (Parameters, ModRedundant…)"* field, insert a blank line before the ModRedundant information (i.e., the line B 116 5157 S 15 -0.1) and two blank lines after the ModRedundant information. Save as a new Gaussian16 input file, with the name **PEP.com**. Launch QM/MM calculation on the terminal window:

- g16 **PEP.com &**

Again, this calculation will take ages. Stop the calculation and copy the file **PEP.log** from the **Files** folder. This would be the result you would get if you wait for a few days.

Open the output file (**PEP.log)** on molUP, select "All optimized structures" option.

Identify the approximate transition state structure and activation energy through the plot available in the "*Output"* tab of molUP. Measure the key interatomic distances of the reaction all along the potential energy profile. Pay particular attention to their values at the reactant and transition state.

Identify the approximate product structure and reaction energy through the plot available in the "*Output"* tab of molUP.

Compare the calculated activation energy with the experimental activation free energy (19.0 kcal/mol). As these reactions are dominated by enthalpy, the energy barrier that you calculate should not be far from the experimental free energy barrier.

In a real research case, more steps would be needed. Namely:

i) To use the transition state structure you have identified as a starting guess to perform a full, unconstrained transition state optimization; ii) to determine the connected reactants and products through internal reaction coordinate calculations; iii) to calculate the zero point energies of all species; iv) to calculate rigid rotor/harmonic oscillator entropies to obtain free energies for all species; v) to recalculate the final energies with higher theoretical levels, or at least, with larger basis sets.

If you want to go through this amazing process up to the end, you can try to finish it by yourself, maybe using some of the author's review papers to help you through this set of calculations.

FINAL REMARKS:

This workshop has given you a basic idea on how to run QM/MM calculations. As any other area of science, you will need years to fully dominate the techniques. Our purpose is to give you a bit of the flavor of this field, to help you to understand better the conferences and seminars that deal with QM/MM calculations, and to help you to progress with QM/MM calculations by yourself. We hope you have enjoyed.

6.6. Outcomes and Future Prospects

In the past decades, with the increasing availability of the internet and Web 2.0 based technologies, access to information changed dramatically. Instead of browsing through magazines such as National Geographic or Discovery for the latest news in science and technology, today's readers might simply use Google or check their social media [278]. Unfortunately, however, access to scientific information from scattered sources is often the cause of misinformation and fake news proliferation.

In tandem, society faces social and structural challenges that require solid scientific judgment to be solved. For example, we have witnessed closely how the advance of genomics, biomedicine, technology, and scientific research has led to effective vaccination strategies one year after identifying COVID-19. This required strong science communication between research scientists, pharmaceutical companies, and policymakers. The role of science in the COVID-19 pandemic is an example of how accurate science communication can effectively change the course of global events.

As doctoral students, it is often our task to communicate our findings and results to a wide range of audiences - from experts in our field, to an online audience and, even, to our family and friends. As a result, we may notice how our speech adapts depending on who is listening. By doing so, we recognize that not everyone is academically equipped to understand the many-step processes and protocols of scientific research.

Throughout the doctoral program in Chemistry, I have contributed in varying extent to four literature reviews, a workshop tutorial, a workbook for the Bioinformatics class, and two articles on the magazine Ciência Elementar aimed at the general public. I believe that it is our duty to communicate science over the course of our careers, to educate fellow citizens. It is only through knowledge that consistent progress is made possible.

Chapter 7. Final Remarks

Enzymes are one of the most promising natural resources available in Nature. Their wide range of substrate selectivity, the vast conditions they operate in, and the high rate-acceleration they achieve through the precise preorganization of active site residues, all of this and more is the result of evolution.

By studying the catalytic mechanism of *Tt*GE, we sought to describe the tailored geometric and electrostatic complementarities between its active site residues and a ligand representative of its natural substrate. Our study described ten stationary points of the *Tt*GE catalytic mechanism, alongside its limiting step and energetic barrier. With a clear depiction of its catalytic mechanism, we wondered: how can we improve its catalytic function? Protein design and engineering have been on the quest to devise strategies to modify, enhance and expand enzymatic activity [279]. By calculating the energetic contribution per residue for the reactant and the first and second transition states of the reactions through the deletion of every single residue within 10 Å of the active-site region, we identified residues that should contribute to the enhancement of the *Tt*GE enzymatic activity. Hence, we believe these are a considerable pool of beneficial mutations that can be further explored through *in vitro* assays.

But successful enzymatic catalysis is not dependent on the catalysis-prone organization of the active site alone. In the case of multi-domain megaenzymes, the level of complexity achieved throughout evolution is extraordinary. Our group has studied the type I animal FAS in-depth, resolving its entire structure, encompassing seven catalytic domains in each of its two monomers. We have also mapped the interactions at the contact interfaces between the tethered ACP substrate-shuttling domain and the other five catalytic domains. This study has served as a scaffold for the subsequent studies on the catalytic mechanism of each of its catalytic domains. Despite our extensive studies and literature survey, many questions are yet to be answered regarding the unusual protein architecture of type I animal FAS. For

instance, what is the explanation for the emergence of an ACP tethered to a multi-enzyme during evolution? How can we further explore the potential of FAS systems for the production of tailored fatty acid molecules?

Another intriguing and vastly unexplored field is that of snake venom. With a growing list of approved drugs derived from proteins identified in snake venom, it is clear that snake venom is such a reservoir for bioactive compounds useful for medicinal purposes. However, one cannot forget that these valuable compounds originated from venom. Snakebite envenoming is a neglected tropical disease with over 100 000 victims every year, particularly in impoverished rural areas in the warmer tropics and subtropics. Thus, we collaborated on the structural, enzymatic, and pharmacological profile of the basic (D49) $sPLA_2$ isolated from the *Agkistrodon piscivorus leucostoma.* Despite the study's many findings, it is necessary to study its catalytic mechanism to devise an antidote for the action of this particular phospholipase A_2.

After the work done on three different subjects, plus the effort to communicate scientific findings to both the academic and general public, we hope to have contributed in a significant way, touching on issues related to climate change, efficient use of waste biomass, snakebite envenoming and investigation of valuable bioactive compounds found in snake venom.

Chapter 8. Appendix

8.1. Evaluation of a flavonoids library for inhibition of pancreatic α-amylase towards a structure–activity relationship

Carina Proença[a], Marisa Freitas[a], Daniela Ribeiro[a], Sara M. Tomé[b], Eduardo F. T. Oliveira[c], Matilde F. Viegas[c], Alberto N. Araújo[a], Maria J. Ramos[c], Artur M. S. Silva[b], Pedro A. Fernandes[c], and Eduarda Fernandes[a]

[a] LAQV, REQUIMTE, Laboratory of Applied Chemistry, Department of Chemical Sciences, Faculty of Pharmacy, University of Porto, Porto, Portugal

[b] Department of Chemistry and QOPNA, University of Aveiro, Aveiro, Portugal

[c] UCIBIO, REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Porto, Portugal

---

Proença, C., Freitas, M., Ribeiro, D., Tomé, S. M., Oliveira, E. F. T., **Viegas, M. F.**, Araújo, A. N., Ramos, M. J., Silva, A. M. S., Fernandes, P. A., & Fernandes, E. (2019). ***Evaluation of a flavonoids library for inhibition of pancreatic α-amylase towards a structure–activity relationship.*** In Journal of Enzyme Inhibition and Medicinal Chemistry 34(1), 577–588. https://doi.10.1080/14756366.2018.1558221

JOURNAL OF ENZYME INHIBITION AND MEDICINAL CHEMISTRY
2019, VOL. 34, NO. 1, 577–588
https://doi.org/10.1080/14756366.2018.1558221

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS Check for updates

Evaluation of a flavonoids library for inhibition of pancreatic α-amylase towards a structure–activity relationship

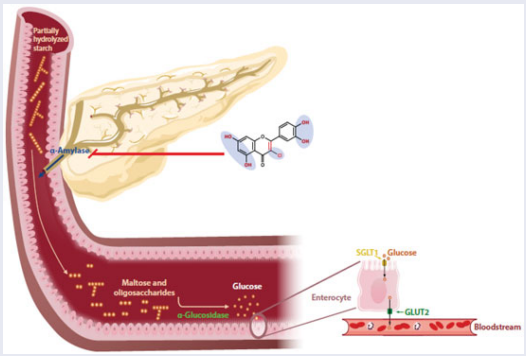
Carina Proença[a], Marisa Freitas[a], Daniela Ribeiro[a], Sara M. Tomé[b], Eduardo F. T. Oliveira[c], Matilde F. Viegas[c], Alberto N. Araújo[a], Maria J. Ramos[c], Artur M. S. Silva[b], Pedro A. Fernandes[c] and Eduarda Fernandes[a]

[a]LAQV, REQUIMTE, Laboratory of Applied Chemistry, Department of Chemical Sciences, Faculty of Pharmacy, University of Porto, Porto, Portugal; [b]Department of Chemistry and QOPNA, University of Aveiro, Aveiro, Portugal; [c]UCIBIO, REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Porto, Portugal

ABSTRACT

α-Amylase has been considered an important therapeutic target for the management of type 2 diabetes *mellitus* (T2DM), decreasing postprandial hyperglycaemia (PPHG). In the present work, a panel of 40 structurally related flavonoids was tested, concerning their ability to inhibit α-amylase activity, using a microanalysis screening system, an inhibitory kinetic analysis and molecular docking calculations. From the obtained results, it was possible to observe that the flavone with a -Cl ion at 3-position of C-ring, an –OH group at 3′- and 4′- positions of B-ring and at 5- and 7- positions of A-ring and the C2 = C3 double bond, was the most active tested flavonoid, through competitive inhibition. In conclusion, some of the tested flavonoids have shown promising inhibition of α-amylase and may be considered as possible alternatives to the modulation of T2DM.

GRAPHICAL ABSTRACT



ARTICLE HISTORY
Received 26 September 2018
Revised 4 December 2018
Accepted 6 December 2018

KEYWORDS
α-Amylase inhibition; diabetes; docking; flavonoids; *in vitro*

Introduction

α-1,4-Glucan-4-glucanohydrolases (EC 3.2.1.1.), belonging to the family 13 of glycoside hydrolases (GH) and found in saliva and pancreatic juice, are known by the common name of α-amylase[1,2]. Both are isozymes expressed respectively from genes AMY1 and AMY2 and are composed of 496 amino acid residues in a single polypeptide chain, presenting a 97% identity of the amino acids sequence and a 92% similarity of the residues in the catalytic domains[3–5]. They have three different domains (A, B and C) and a common calcium-binding site that stabilizes the interface between the central A domain and the variable B domain[6,7]. The catalytic triad (Asp, Asp, Glu) is present in the A domain and the active site is located in a long cleft between the carboxyl end of both A and B domains. The C domain consists of a *β*-sheet structure linked to the A domain by a simple polypeptide chain[5].

CONTACT Eduarda Fernandes egracas@ff.up.pt LAQV, REQUIMTE, Laboratory of Applied Chemistry, Department of Chemical Sciences, Faculty of Pharmacy, University of Porto, Rua de Jorge Viterbo Ferreira no. 228, 4050-313 Porto, Portugal; Pedro A. Fernandes pafernan@fc.up.pt UCIBIO, REQUIMTE, Department of Chemistry and Biochemistry, Rua do Campo Alegre, w/n, 4169-007 Porto, Portugal

Supplemental data for this article can be accessed here.

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

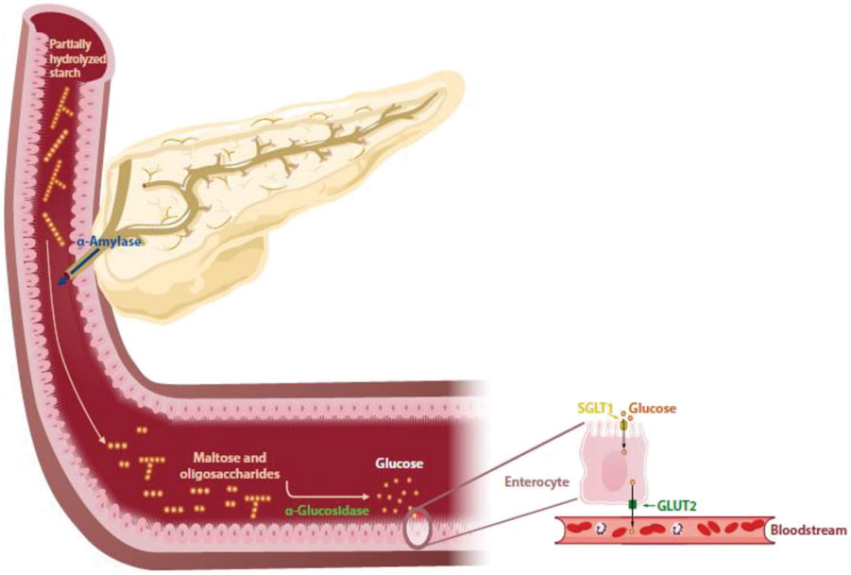


Figure 1. Schematic representation of the pancreatic α-amylase and α-glucosidase activity. After a meal, α-amylase synthesized in pancreas and released in the duodenum, catalyzes the hydrolysis of α-1,4 glycosidic linkages in partially hydrolyzed starch (amylopectin and amylose). From this reaction, intermediate unbranched, such as maltose and maltotriose, and branched (α-limit dextrins) oligosaccharides are formed. α-Glucosidase present in the brush border of the intestinal epithelium (enterocytes) is responsible for the final step of carbohydrates digestion, prior to their absorption. This enzyme converts the disaccharides and oligosaccharides into glucose, which is then transported by sodium/glucose co-transporter 1 (SGLT1) from the intestinal lumen to the cytosol of enterocytes. In turn, glucose transporter 2 (GLUT2), found in the basolateral membrane of enterocytes, transports glucose from cytosol to blood via facilitated diffusion.

α-Amylase catalyzes the initial hydrolysis of starch and other carbohydrate polymers into shorter oligosaccharides through cleavage of α-1,4- bonds[5,7,8]. The salivary isozyme provides an initial partial cleavage into shorter oligomers (10–30%)[7]. Once these partially digested saccharides reach the gut, they are extensively hydrolyzed by the α-amylase synthesized in the pancreas and excreted in the lumen into smaller oligosaccharides, such as maltose, maltotriose and α-limit dextrins (Figure 1)[3]. These sugars are finally broken down into glucose by the intestinal brush border α-glucosidases, which is in turn absorbed from the intestinal mucosa into the portal blood, by means of the glucose transporter (GLUT2) and sodium-glucose co-transporter 1 (SGLT1), leading to postprandial hyperglycaemia (PPHG)[2]. Impaired regulation of PPHG constitutes a common feature in type 2 diabetes *mellitus* (T2DM), the most prevalent form of diabetes, accounting for about 90% of all diabetes cases[2,9,10].

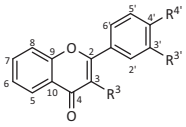
The pancreatic α-amylase activity has been targeted for inhibition by means of the so-called starch blockers in order to mitigate PPHG[1,3]. Acarbose is the most widely prescribed α-amylase inhibitor, and in spite of its efficiency in the control of PPHG, the administration of this drug is associated with gastrointestinal adverse effects in diabetic patients, namely abdominal distention, flatulence and diarrhoea[11]. Thus, the search and development of new effective and safer agents, able to control glucose levels is of high importance for the management of T2DM[12].

In the last few years, the activity of flavonoids has been studied even up to clinical trials concerning the modulation of diabetes in humans[13,14]. Promising inhibitory potential against important targets related with the T2DM pathophysiology was evidenced regarding the inhibition of isolated enzymes such as α-glucosidase[15] and protein tyrosine phosphatase 1B (PTP1B)[16]. Some bibliographic reviews compile the obtained results of the antidiabetic activity of flavonoids[17–20]. Further isolated studies also describe flavonoids as α-amylase inhibitors, such as apigenin[21,22], luteolin[21,23], kaempferol[21,22], naringenin[21], quercetin[21,22,24,25], myricetin[21,22,25], chrysin[22] and baicalein[21,22]. Nevertheless, significant differences in the experimental conditions among studies, such as the source and concentrations of enzyme and substrate and the different incubation times applied, turn difficult the establishment of a reliable structure-activity relationship. To fill this gap, a panel of 40 structurally related flavonoids (Figure 2), most of them studied for the first time, was assessed regarding their inhibitory activity against α-amylase. This study comprehends an *in vitro* microanalysis screening system, modelling of kinetics inhibition and an *in silico* molecular docking analysis.

Materials and methods

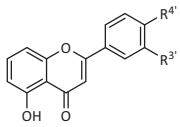
Chemicals

α-Amylase from porcine pancreas, 2-chloro-4-nitrophenyl-α-D-maltotrioside (CNPG3), **acarbose**, DMSO, $NaHPO_4$, Na_2HPO_4, flavonoids **D3** (baicalein), **D4** (apigenin), **D6** (kaempferol), **D7** (luteolin), **D8** (quercetin), **D9** (myricetin), **D10** (morin), **D17** (acacetin), **D19** (rutin), **E1** (naringenin), **E2** (eriodictyol) and **E3** (taxifolin) were obtained from Sigma-Aldrich Co. LLC (St. Louis, MO). The following flavonoids were obtained from Indofine Chemical Company, Inc. (Hillsborough, NJ): **A1**, **A2**, **A3**, **A4**, **A5**, **A6**, **B1**, **B2**, **B3**, **B4**, **C1**, **C2**, **C3**, **C4**, **C5**, **C6**, **C7**, **C8**, **D1** (chrysin), **D2** (galangin) and **D5**. The



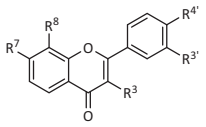
A1-6

A1: R^3= $R^{3'}$= $R^{4'}$= H (Flavone)
A2: R^3= H, $R^{3'}$= OH, $R^{4'}$= H
A3: R^3= $R^{3'}$= H, $R^{4'}$= OH
A4: R^3= H, $R^{3'}$= $R^{4'}$= OH
A5: R^3= $R^{3'}$= $R^{4'}$= OH
A6: R^3= OH, $R^{3'}$= $R^{4'}$= OMe



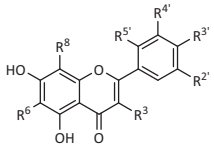
B1-4

B1: $R^{3'}$= $R^{4'}$= H
B2: $R^{3'}$= OH, $R^{4'}$= H
B3: $R^{3'}$= H, $R^{4'}$= OH
B4: $R^{3'}$= $R^{4'}$= OH



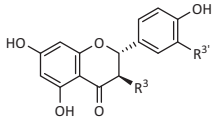
C1-8

C1: R^3= $R^{3'}$= $R^{4'}$= H, R^7= OH, R^8= H
C2: R^3= H, $R^{3'}$= OH, $R^{4'}$= H, R^7= OH, R^8= H
C3: R^3= $R^{3'}$= H, $R^{4'}$= R^7= OH, R^8= H
C4: R^3= $R^{3'}$= OH, $R^{4'}$= H, R^7= OH, R^8= H
C5: R^3= OH, $R^{3'}$= H, $R^{4'}$= R^7= OH, R^8= H
C6: R^3= H, $R^{3'}$= $R^{4'}$= R^7= OH, R^8= H
C7: R^3= OH, $R^{3'}$= OMe, $R^{4'}$= H, R^7= OMe, R^8= H
C8: R^3= OH, $R^{3'}$= H, $R^{4'}$= R^7= OMe, R^8= H



D1-19

D1: $R^{2'}$= R^3= $R^{3'}$= $R^{4'}$= $R^{5'}$= R^6= R^8= H (Chrysin)
D2: $R^{2'}$= H, R^3= OH, $R^{3'}$= $R^{4'}$= $R^{5'}$= R^6= R^8= H (Galangin)
D3: $R^{2'}$= R^3= $R^{3'}$= $R^{4'}$= H, $R^{5'}$= R^6= OH, R^8= H (Baicalein)
D4: $R^{2'}$= R^3= H, $R^{3'}$= OH, $R^{4'}$= $R^{5'}$= R^6= R^8= H
D5: $R^{2'}$= R^3= $R^{3'}$= H, $R^{4'}$= OH, $R^{5'}$= R^6= R^8= H (Apigenin)
D6: $R^{2'}$= H, R^3= OH, $R^{3'}$= H, $R^{4'}$= OH, $R^{5'}$= R^6= R^8= H (Kaempferol)
D7: $R^{2'}$= R^3= H, $R^{3'}$= $R^{4'}$= OH, $R^{5'}$= R^6= R^8= H (Luteolin)
D8: $R^{2'}$= H, R^3= $R^{3'}$= $R^{4'}$= OH, $R^{5'}$= R^6= R^8= H (Quercetin)
D9: $R^{2'}$= R^3= OH, $R^{3'}$= H, $R^{4'}$= OH, $R^{5'}$= R^6= R^8= H (Morin)
D10: $R^{2'}$= H, R^3= Cl, $R^{3'}$= $R^{4'}$= OH, $R^{5'}$= R^6= R^8= H
D11: $R^{2'}$= R^3= H, $R^{3'}$= $R^{4'}$= OH, $R^{5'}$= H, R^6= Cl, R^8= H
D12: $R^{2'}$= R^3= H, $R^{3'}$= $R^{4'}$= OH, $R^{5'}$= R^6= H, R^8= Cl
D13: $R^{2'}$= H, R^3= Cl, $R^{3'}$= $R^{4'}$= OH, $R^{5'}$= R^6= H, R^8= Cl
D14: $R^{2'}$= R^3= H, $R^{3'}$= $R^{4'}$= OH, $R^{5'}$= H, R^6= R^8= Cl
D15: $R^{2'}$= R^3= H, $R^{3'}$= OMe, $R^{4'}$= OH, $R^{5'}$= R^6= R^8= H (Chrysoeriol)
D16: $R^{2'}$= R^3= $R^{3'}$= H, $R^{4'}$= OMe, $R^{5'}$= R^6= R^8= H (Acacetin)
D17: $R^{2'}$= R^3= $R^{3'}$= H, $R^{4'}$= OH, $R^{5'}$= H, R^6= OMe, $R^{8'}$= H (Hispidulin)
D18: $R^{2'}$= R^3= H, $R^{3'}$= $R^{4'}$= OMe, $R^{5'}$= R^6= R^8= H
D19: $R^{2'}$= H, R^3= rutinose, $R^{3'}$= $R^{4'}$= OH, $R^{5'}$= R^6= R^8= H (Rutin)



E1-3

E1: R^3= $R^{3'}$= H (Naringenin)
E2: R^3= H, $R^{3'}$= OH (Eriodictyol)
E3: R^3= $R^{3'}$= OH (Taxifolin)

Figure 2. Chemical structures of the studied flavonoids.

flavonoids **D11**, **D12**, **D13**, **D14**, **D15, D16**and **D18** were synthesized according to previous companion papers[26,27].

In vitro *pancreatic α-amylase inhibition assay*

The evaluation of inhibitory effect on α-amylase activity was based on the method described by Trinh et al.[28], with slight modifications. In each assay, the α-amylase mediated hydrolysis of the substrate CNPG3 into 2-chloro-nitrophenol (CNP), 2-chloro-4-nitrophenyl-α-D-maltoside (CNPG2), maltotriose and glucose was monitored. The initial rate of CNP generation, measured spectrophotometrically at 405 nm, is proportional to the concentration of α-amylase present. In brief, in a 96-well plate, the enzyme (0.2 U/mL), dissolved in 20 mM phosphate buffer (20 mM Na_2HPO_4

and 7 mM NaCl, pH 6.8) was exposed to the flavonoids under study (0–200 μM), dissolved in DMSO [final concentration of DMSO of 4.5% (*v/v*)]. After an incubation time of 10 min, at 37 °C, the reaction was started by the addition of CNPG3 (0.5 mM), dissolved in 20 mM phosphate buffer (pH 6.8) and monitored along 30 min, at 37 °C, in a microplate reader (Synergy HT, BIO-TEK), with the wavelength set to 405 nm. The concentration of α-amylase and the tested flavonoids was determined before the addition of the substrate. Increasing absorbance values obtained within the 5 to 30 min interval (slope) served for the calculation of catalytic rates and the results in the form of inhibition percentages, representing at least three independent experiments. **Acarbose** (0–16 μM) was used as positive control. The amount of DMSO used did not interfere with the assay. The results of the *in vitro* inhibitory activity of flavonoids against the pancreatic α-amylase activity are expressed as mean ± standard error of mean (SEM). Statistical comparison between the active flavonoids was estimated by applying the one-way analysis of variance (ANOVA). Differences were considered to be significant at p values lower than 0.05. All the statistical analysis were performed using GraphPad Prism™ (version 7.0; GraphPad Software).

Inhibitory kinetics of α-amylase

The estimation of kinetic parameters and prediction of the actual mechanism of inhibition was performed by means of Microsoft Office Excel™ spreadsheets and using the Solver supplement add-in, according to Bezerra et al.[29] and Dias et al.[30]. Therefore, the kinetics of conversion of CNPG3 by α-amylase in each microplate well was fitted by nonlinear least squares regression[31] using the general model for enzymatic reactions translated by the Equation (1), and each one of its simplifications regarding the different types of inhibition:

$$v_{inic} = \frac{V_{máx}(S)}{K_m\left(1+\frac{I}{K_{ic}}\right)+(S)\left(1+\frac{I}{K_{iu}}\right)} \quad (1)$$

where v_{inic} and V_{max} represent respectively the initial velocity of formation of absorbing CNP in μmol/min and the maximum achievable velocity when for the 0.2 U/mL of enzyme used all catalytic sites are saturated by the substrate, S is the CNPG3 concentration in mM, K_m is the Michaelis–Menten constant in mM and K_{ic}, K_{iu} are the constants for competitive and uncompetitive inhibition processes expressed in μM^{-1}.

To study the enzyme kinetics, nonlinear regression was applied to the nontransformed data for the most active flavonoids of each group: **B4** (0–200 μM), **C5** (0–100 μM), **D11** (0–100 μM) and **acarbose** (0–2 μM). For each concentration of inhibitor, three concentrations of the substrate CNPG3 were tested: 0.25, 0.5 and 1 mM. The obtained results represent at least three independent experiments.

For each tested condition, the best guesses for equation parameters were thrown by Solver tool after iterative minimization of the sum of squared residuals between the experimental values of V_{max} and the corresponding counterparts estimated according to each model used. Hence, the values obtained for the parameters of the simplest model (without inhibition) were used as initial values, proceeding to competitive inhibition, noncompetitive inhibition, uncompetitive inhibition and finishing with the more complex model, mixed inhibition. The actual mechanism of inhibition was established after comparison and discrimination between the models with a different number of parameters applying the extra sum-of-square F test[32] and the Akaike information criterion (AIC) test[33].

The jackknife procedure was applied to determinate the error of the kinetic constant values. It consisted of the calculation of standard deviation of all estimates guessed by Solver when each experimental data point was in turn drawn from the initial set[34].

Molecular docking analysis

Seventeen entries in the Protein Data Bank (PDB) correspond to *Sus Scrofa* α-amylase (E.C. 3.2.1.1.). From these, it was chosen the PDB entry 1HX0[35] corresponding to the highest resolution X-ray structure (1.38). This structure was co-crystalized with a trisaccharide inhibitor K2, derived from acarbose after α-amylase cleaved one glucose unit at the reducing end. The crystal structure shown two α-1,4-linked K2 molecules, condensed *in situ*. The K2 unit bound at the active site was used to compare the binding mode of the tested flavonoids and as a template to model the substrate for receptor minimization and conformational sampling.

The protonation states of the enzyme were predicted using the Protoss server[36]. Accordingly, Glu233 and Glu390 were modelled in the neutral form and His101, His201 and His215 were modelled in the positively charged for, with both side chain nitrogen atoms protonated. All other residues were protonated according to their pKa, at pH = 7.

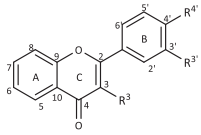
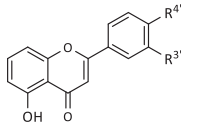
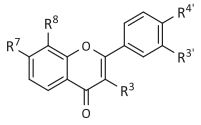
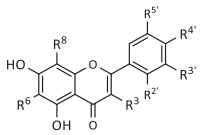
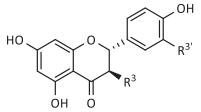
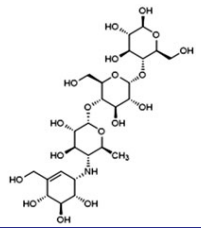
From the hexasaccharide compound present in 1HXO (resultant of two α-1,4-linked K2 molecules), the substrate maltohexose was modelled. On this system, Molecular Mechanics minimization was performed, followed by 20 ns of classical molecular dynamics (MD) simulation to sample receptor conformations. This protocol was performed within the AMBER 12 suite[37] using the ff99SB[38] force field for the protein and GLYCAM06[39] force field for maltohexose. Further details for the energy minimization and MD simulation can be found in Supporting Information.

To choose and validate the conformation of the receptor for the docking protocol several receptor conformations were evaluated: the X-ray conformation, the conformation of the receptor minimized with maltohexose and four representative conformations of the clustered MD simulation. The suitability of each conformation was assessed by evaluating if the docking of the inhibitor acarbose in the active site reproduced the X-ray conformation and how well the ranking of the tested flavonoids with the lowest IC_{50} was reproduced. The conformation of the receptor with the best performance was obtained after 3.3 ns of simulation.

The binding mode of the selected flavonoids was predicted by means of the AutoDock Vina[40]. The binding poses were generated using default parameters (see Supporting Information for details), such as the exhaustiveness of the global search (8) or the maximum number of output poses (9). The search space was confined to a box of 19 Å × 12 Å × 17 Å centered on the site occupied by the glycosidic oxygen between the −1 and −2 glucose units of maltohexose. The receptor was treated as a rigid body and prepared using openbabel[41] for atom-typing and charge assignment using Gasteiger–Marsili charges[42]. The atom-types and charges for the tested flavonoids were attributed the same way.

Subsequently, the poses predicted by AutoDock Vina for each flavonoid were scored using the scoring function GoldScore within the GOLD software suite[43], using the standard parameters for a "Rescoring Run" with local optimization.

Table 1. Inhibitory activity of flavonoids against porcine pancreatic α-amylase (IC_{50}, μM ± SEM).

Compound	Structure	$R^{2'}$	R^3	$R^{3'}$	$R^{4'}$	$R^{5'}$	R^6	R^7	R^8	IC_{50} (μM) or Inhibition (%)
A1 (Flavone)		–	H	H	H	–	–	–	–	$<20\%^{200}$ μM*
A2		–	H	OH	H	–	–	–	–	$<20\%^{200}$ μM*
A3		–	H	H	OH	–	–	–	–	$<20\%^{200}$ μM*
A4		–	H	OH	OH	–	–	–	–	$46 \pm 2\%^{200}$ μM*
A5		–	OH	OH	OH	–	–	–	–	$31 \pm 3\%^{200}$ μM*
A6		–	OH	OMe	OMe	–	–	–	–	$<20\%^{200}$ μM*
B1		–	–	H	H	–	–	–	–	$<2\%^{200}$ μM*
B2		–	–	OH	H	–	–	–	–	$<20\%^{200}$ μM*
B3		–	–	H	OH	–	–	–	–	$<20\%^{200}$ μM*
B4		–	–	OH	OH	–	–	–	–	148 ± 5 μM[a,h]
C1		–	H	H	H	–	–	OH	H	$<20\%^{200}$μM*
C2		–	H	OH	H	–	–	OH	H	$28 \pm 3\%^{200}$μM*
C3		–	H	H	OH	–	–	OH	H	$36 \pm 3\%^{200}$μM*
C4		–	OH	OH	H	–	–	OH	H	$34 \pm 3\%^{200}$μM*
C5		–	OH	H	OH	–	–	OH	H	59 ± 4 μM[b,c]
C6		–	H	OH	OH	–	–	OH	H	131 ± 1 μM[a,d]
C7		–	OH	OMe	H	–	–	OMe	H	$< 20\ \%^{200}$ μM*
C8		–	OH	H	OMe	–	–	OMe	H	$< 20\%^{100}$ μM*
D1 (Chrysin)		H	H	H	H	H	H	–	H	$22 \pm 3\%^{200}$μM*
D2 (Galangin)		H	OH	H	H	H	H	–	H	$< 20\%^{200}$ μM*
D3 (Baicalein)		H	H	H	H	H	OH	–	H	$< 20\ \%^{200}$μM*
D4		H	H	OH	H	H	H	–	H	≈ 200 μM[e]
D5 (Apigenin)		H	H	H	OH	H	H	–	H	122 ± 7 μM[a,f]
D6 (Kaempferol)		H	OH	H	OH	H	H	–	H	118 ± 7 μM[a,f]
D7 (Luteolin)		H	H	OH	OH	H	H	–	H	78 ± 3 μM[b,g]
D8 (Quercetin)		H	OH	OH	OH	H	H	–	H	138 ± 5 μM[a]
D9 (Myricetin)		H	OH	OH	OH	OH	H	–	H	107 ± 6 μM[d,f]
D10 (Morin)		OH	OH	H	OH	H	H	–	H	$23 \pm 1\ \%^{200}$μM*
D11		H	Cl	OH	OH	H	H	–	H	44 ± 3 μM[c]
D12		H	H	OH	OH	H	Cl	–	H	≈ 200 μM[e]
D13		H	H	OH	OH	H	H	–	Cl	173 ± 5 μM[e,h]
D14		H	Cl	OH	OH	H	H	–	Cl	82 ± 4 μM[g]
D15		H	H	OH	OH	H	Cl	–	Cl	98 ± 3 μM[f]
D16 (Chrysoeriol)		H	H	OMe	OH	H	H	–	H	192 ± 7 μM[e]
D17 (Acacetin)		H	H	H	OMe	H	H	–	H	$<20\%^{200}$ μM*
D18		H	H	OMe	OMe	H	H	–	H	$33 \pm 3\ \%^{200}$ μM*
D19 (Rutin)		H	Rutinose	OH	OH	H	H	–	H	$<20\%^{200}$ μM*
E1 (Naringenin)		–	H	H	–	–	–	–	–	$<20\%^{200}$ μM*
E2 (Eriodictyol)		–	H	OH	–	–	–	–	–	$<20\%^{200}$ μM*
E3 (Taxifolin)		–	OH	OH	–	–	–	–	–	$<20\%^{200}$ μM*
Positive control: Acarbose		–	–	–	–	–	–	–	–	1.3 ± 0.2 μM[i]

*Inhibition (mean, % ± SEM) at the highest tested concentration (in superscript), at the assay conditions.
The IC_{50} with different lowercase superscript letters are significantly different from each other ($p < 0.05$).

Results

In vitro pancreatic α-amylase inhibition assay

The tested flavonoids were divided into five groups (A–E), according to their substitution pattern. The substitution pattern of flavonoids from the A group differs at 3-position of the C-ring and at 3′- and 4′-positions of the B-ring; flavones from B group have an –OH group at 5-position of the A-ring and different substitutions at 3′- and 4′-positions of the B-ring; flavonoids from C group show –OH, –OMe and/or –OBn groups at 3-position of the C-ring, at 3′- and 4′-positions of the B-ring and at 7- and 8-position of the A-ring; flavonoids from D group present –OH substitutions at 5- and 7- positions of the A-ring and also present –OH, –OMe and/or –Cl substitutions at 3-position of the C-ring, at 2′-, 3′-, 4′- and 5′-positions of the B-ring and at 6- and 8-positions of the A-ring; flavanones from E group have –OH groups at 5- and 7-positions of the A-ring and their substitution pattern differs in the presence or absence of –OH groups at 3′- and 4′-positions of the B-ring (Table 1).

No relevant inhibitory activities were observed, up to the highest tested concentration (200 μM), for the flavones belonging to A group with different substitutions in 3-position of C-ring and in 3′- and 4′-positions of B-ring.

For flavones of B group, with an –OH group at 5-position of A-ring and different substitutions in 3′- and 4′-positions of B-ring, the most effective was **B4,** presenting an IC_{50} value of 148 ± 5 μM. This result demonstrates that the presence of the catechol group at B-ring translated in a relatively slight increase of the inhibitory effect over α-amylase catalysis when compared to the other flavones of this group. The presence of an –OH group at 5-position of A-ring also seems to contribute to such effect, since flavone **B4** was more active than flavone **A4** (also with 3′- and 4′–OH in B-ring, but without 5–OH in A-ring).

From C group, which includes flavonoids with different substitutions at 3-position of C-ring, at 3′ and 4′ positions of B-ring and at 7 and 8 positions of A-ring, the most active flavonoid was compound **C5**, tested here for the first time, with an IC_{50} value of 59 ± 4 μM. It seems that the presence of –OH groups at 3-position of C-ring, at 4′-position of B-ring and at 7-position of A-ring, is accompanied by a significant inhibitory effect on α-amylase. On the other hand, the presence of -OMe groups in the flavonoid scaffold (**C7**, **C8**) did not bring any advantage for the intended effect, since the initial conversion rates of the CNPG3 substrates remained unchanged even for the highest tested concentration (200 μM).

All members of D group have –OH groups at 5- and 7-positions of A-ring and the substitution pattern varies according to the presence or absence of –Cl ions, –OH and –OMe groups in their structure. The most effective flavonoid found was the chlorinated flavone **D11**, with an IC_{50} value of 44 ± 3 μM. Thus, the presence of a –Cl ion at 3-position of C-ring, as well as the presence of a catechol in B-ring and –OH groups at 5- and 7-positions of A-ring, showed to be crucial for the inhibition of α-amylase activity. Comparing flavonoid **D11** with **D8** (quercetin) (IC_{50}= 138 ± 5 μM), it is clear that the substitution of the Cl- ion for an OH-group at 3-position of C-ring results in a significant decrease of the corresponding inhibitory activity. It was also observed that the presence of a -OMe group at 3′- position of B-ring [**D16** (chrysoeriol)] slightly improves the ability to inhibit α-amylase activity (IC_{50}= 192 ± 7 μM), comparing with **D17** (acacetin), which presents a -OMe group at 4′-position of B-ring. An increase of the number of -OMe groups in the structure of flavonoids is not favourable for the intended effect, as it can be concluded from a comparison of **D16** (chrysoeriol) with **D18**, the later with an extra -OMe group at 4′-position of B-ring. A glycosylated flavonoid, **D19** (rutin), was also tested, but no activity was found up to the highest tested concentration (200 μM).

The studied flavanones of E group, **E1** (naringenin), **E2** (eriodictyol) and **E3** (taxifolin) were not able to inhibit α-amylase activity, at the maximum tested concentration (200 μM). These results demonstrate that the presence of C2 = C3 double bond is very relevant for the intended effect.

Table 2. Type of inhibition and V_{max}, K_m, K_{ic}, and K_{iu}values ± SD (μM) for the inhibition of α-amylase by the selected flavonoids.

Compound	Type of inhibition	V_{max} (μmol/min)	K_m (mM)	K_{ic} (μM)	K_{iu} (μM)
B4	Competitive	12.0 ± 0.6	1.2 ± 0.1	132 ± 15	–
C5	Competitive	10.5 ± 0.9	1.0 ± 0.1	71 ± 7	–
D11	Competitive	15 ± 4	1.1 ± 0.4	21 ± 3	–
Positive control:					
Acarbose	Mixed	12.0 ± 0.9	0.9 ± 0.1	6 ± 3	0.71 ± 0.09

Inhibitory kinetics of α-amylase

In accordance with the described results, the most effective flavonoids of B, C and D groups, **B4**, **C5** and **D11**, respectively, and the positive control, **acarbose**, were henceforth selected for their type of inhibition calculation. No flavonoids from A and E group were tested, due to their lack of inhibitory activity.

By comparing the sum of the square errors values among the different enzyme inhibition models and, for models with different number of parameters, applying the extra sum-of-square *F* test and the AIC test, it was possible to select the best model (without inhibition, competitive inhibition, noncompetitive inhibition, uncompetitive inhibition or mixed inhibition). In that sense, the model with the lowest sum of the square error value corresponds to the best kinetic model. The comparison between models with a different number of parameters can be done by comparing the extra sum-of-square *F* test, at the desired level of probability (f 0.95): if the $-f$ value is superior to 0, the more complex model should be applied. The Akaike test was also used to discriminate the different enzyme inhibition models, with a different number of parameters: the model with the lower corrected AIC (AIC_c) score is the model more likely to be correct.

By the obtained results, it was possible to conclude that **acarbose** [see Supporting Information, Figures 1(S)–5(S)] is a mixed-type inhibitor, and flavonoids **B4** [see Supporting Information, Figures 6(S)–10(S)], **C5** [see Supporting Information, Figures 11(S)–15(S)] and **D11** are competitive inhibitors [see Supporting Information, Figures 16(S)–20(S)].

After the definition of the best model, it was possible to summarize the kinetic constants values (V_{max}, K_m, K_{ic} and/or K_{iu}) (Table 2). The tested compounds presented the following order of K_{ic} values: **B4**>**C5**>**D11**>**acarbose**.

Molecular docking analysis

Binding mode of the selected inhibitors

Once evidenced the inhibition mechanism through experimental data and calculations using Solver supplement of Microsoft Excel Office™, it was predicted the binding pose of flavonoids within the active site of α-amylase. Special attention is given to **B4**, **C5** and **D11** derivatives, the ones exhibiting a higher inhibitory effect in their groups.

The binding pocket of α-amylase is usually divided in –1 to N subpockets, each one creating a particular environment around each sugar ring. The innermost subpocket is numbered as –1

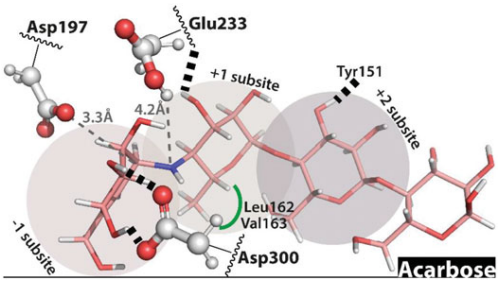


Figure 3. Predicted binding mode for **acarbose**. Hydrogen bonds are represented by dashed lines and hydrophobic interactions by green lines. The distances between the catalytic residues Asp197 and Glu233 and the anomeric carbon/N-glycosidic bond are represented by dashed gray lines.

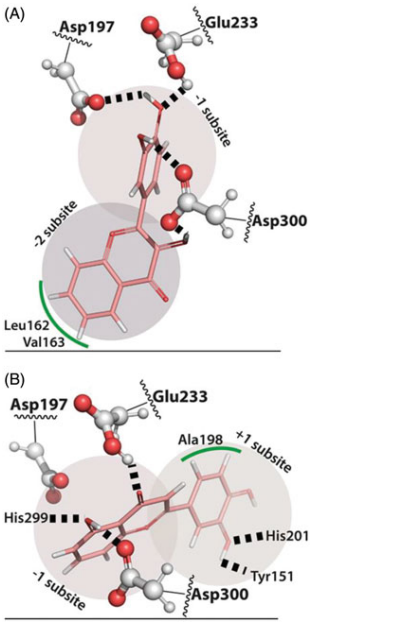


Figure 4. (A) Predicted binding mode of flavonoid **A5**, occupying the –2 and –1 subsites with the more hydrophobic A-ring on the –2 subsite interacting with Leu162 and Val163. (B) The predicted binding mode for **B4** shows the C-ring carbonyl and A-ring hydroxyl hydrogen bonded to the catalytic residues Glu233, Asp300 and His299.

subpocket and is immediately followed the +1 subpocket while the glycosidic bond to be cleaved stays positioned in-between the sugars and also in-between the –1 and +1 subpockets. The next subpockets are numbered +2, +3, etc. An even deeper subpocket (–2) can be occupied by inhibitors, but generally not by natural substrates.

We started by evaluating the accuracy of our Molecular Docking Protocol. For this purpose, we performed a docking assay using the crystal structure of alpha-amylase complexed with acarbose (*Bacillus subtilis,* PDB structure 1UA7, and resolution 2.21 Å). We deleted the acarbose hexasaccharide from the 1UA7 PDB file and re-docked it into the active site, following the protocol described in the methods, using the same parameters, except for the size of the box (which was, in this case, 21.75 × 22.50 × 22.50 Å). The top-ranking solution of the docking assay displayed a very similar pose to that of acarbose in the crystallographic structure [see Supporting Information, Figure 21(S)]. Root mean square deviation (RMSD) analysis of the heavy atoms in the docking solution concurs with visual inspection, with a very low heavy-atom RMSD value of 0.482 Å, confirming that the used protocol has significant predictive power.

The study started with the prediction of the binding mode of acarbose, which in the chosen receptor perfectly reproduces the binding interactions of the substrate maltohexose seen in the MD simulation (Figure 3). This was expected, as it was a condition that we imposed to choose the receptor conformation for the docking calculations. The hydroxymethyl conduritol ring of acarbose accommodates to the acarviosine moiety bound to the –1 and +1 subpockets while the two farthest glucose units extend from the +2 subpocket to onwards. In such conformation, this non-hydrolyzable *N*-glycosidic bond stays moderately distant from the general acid Glu233 (4.2 Å), but the nucleophilic Asp197 is only at 3.3 Å from the anomeric carbon, Asp300 is hydrogen bonded to the 2-hydroxyl and 3-hydroxyl groups of the glucose unit of the –1 subpocket. An additional hydrogen bond is established between the hydroxymethyl group and Asp197.

The most relevant interaction between the glucose unit in the +1 subpocket and the receptor seems to be the hydrophobic interaction between its methyl group and a hydrophobic pocket formed by Leu162 and Val163, and a hydrogen bond between the 3–OH and the backbone of Glu233. For the glucose unit at the +2 subpocket, the most relevant interaction seems to be a hydrogen bond between the 3–OH group and the side-chain of Tyr151. The glucose unit at the +3 subpocket is completely exposed to the solvent and does not interact meaningfully with the receptor.

Among flavonoids from A group, which carry substitutions at the 3-position of C-ring and at 3′- and 4′-positions of B-ring, a clear trend is observed: from flavonoids **A1** to **A5**, the –OH groups at 3-position of C-ring and at 3′-position of B-ring establish hydrogen bonds with Asp300, while the –OH group at 4′-position of B-ring establishes hydrogen bonds with Asp197, with this hydrogen bonding network being optimized in flavonoid **A5**.

Therefore, flavonoids **A1** to **A5** bind the active site with the B-ring oriented towards the catalytic triad (Asp197, Asp300, Glu233), occupying a site equivalent to the –1 subpocket of the K2 inhibitor in the PDB structure 1HXO[35], while the A- and C-rings are pointing outwards occupying the binding site cleft equivalent to the –2 subpocket, which is a pocket not usually occupied by the natural substrate, located in the opposite direction in relation to the +1 subpocket.

For the less substituted **A1** to **A3** flavonoids, the described set of hydrogen bond interactions is not optimized and as such, they do not bind so deep in the active site. Instead, aromatic stacking interactions are observed between the A- and C-rings and Trp58, and between the B-ring and Tyr62.

When two –OH groups are added to B-ring (flavonoids **A4** and **A5**) the hydrogen bonds with the catalytic triad are optimized and these flavonoids bind deeper in the active site, hydrogen bonding with Asp197, Glu233 and Asp300, which might explain the higher inhibitory activity observed for these two flavones within A group [Figure 4(A)].

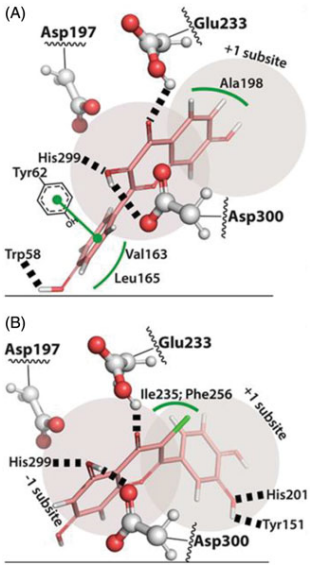


Figure 5. (A) The binding mode predicted for **C5** shows the C-ring carbonyl and –OH group interacting with the active site residues in a similar way as in **B4**. This requires the B-ring to be placed in the −1/−2 subsites, stabilized mostly by hydrophobic interactions with Tyr62, Val163 and Leu165. (B)The best tested inhibitor, flavonoid **D11**, and flavonoid **B4** binds very similarly to α-amylase. The most important difference seems to come from the 3–Cl substitution that interacts with Ile235 and Phe256.

The steric hindrance introduced by methoxylation at 3′- and 4′-positions of B-ring (flavonoid **A6**), flips the binding pose, placing instead the A- and C-rings in the −1 subpocket. This allows the –OH group at 3-position of C-ring to establish a hydrogen bond with Asp300 and Glu233.

As evidenced before, within B group of structurally related flavonoids, the best inhibitor is **B4** with an IC_{50} value of 148 ± 5 μM. This compound shares with **A4** and **A5** the same hydroxyl substituents in the B-ring, but an additional –OH group at 5-position of the A-ring, which jointly promotes a stabilized conformation and an increase of the inhibition of α-amylase activity.

Looking at the pose predictions, it is clear that the C-ring carbonyl and the –OH group at the A-ring, promote the positioning of the A- and C-rings close to the catalytic triad. The orientation of the flavonoid, however, depends on the substitutions present in the B-ring [Figure 4(B)].

The two –OH substitutions present in the B-ring of flavonoid **B4** increase the hydrophilicity of the catechol, which seems to distinguish **B4** from the rest of the group. This allows the B-ring to occupy the +1 subpocket, with His201 hydrogen bonding the oxygen of the –OH group at 3′ position of B-ring, while that same –OH group makes a hydrogen bond with the oxygen of the Tyr151 hydroxyl [Figure 4(B)]. This orientation is unlike **B1** and **B2**, where the more hydrophobic B-ring binds the –2 subpockets, in the more hydrophobic region formed by Val163 and Leu165. Despite positioning the A- and C-ring between the catalytic triad, as the other flavonoids, this different orientation of the A- and C-ring of **B4**, seems to optimize the hydrogen-bonding network around the active site. Namely, the –OH group in the 5-position of A-ring establishes a hydrogen bond with Asp300 and to His299, while the C-ring carbonyl is hydrogen bonded to Glu233.

In turn, from C group, the 3–OH substitution present in C-ring of flavonoids **C4** and **C5** gives rise to a substructural motif of adjoined hydroxyl and carbonyl groups, similarly to the one found in the B group. On the latter, a C-ring carbonyl group adjoined by a –OH group at 4-position of A-ring allows the positioning of A- and C-rings within the catalytic triad. Therefore, the C-ring carbonyl and 3–OH substitution in **C4** and **C5** flavonoids also promotes the placing of the A- and C-rings within the catalytic triad, with the hydrogen of the –OH group at 3-position of C-ring bonded to Asp300 and to His299, and with the Glu233 hydrogen bonding the carbonyl group [Figure 5(A)]. These positions of the carbonyl and –OH groups, exactly as in **B4,** and the same hydrogen-bonding network, are observed between these groups and the active site. This also means that **C4** and **C5** are bound in a flipped orientation when compared to flavonoid **B4**. Thus, these flavonoids occupy the subpockets +2 to −1, with the –OH group at 7-position of A-ring in the −1 subpocket, while the B-ring binds in the other side of the active site cleft, with the B-ring facing the more hydrophobic side of the substrate cleft characterized by Val163, Leu165 and Tyr62 (+2 subpocket). The B-ring establishes aromatic stacking interactions with Tyr62 and in the case of flavonoid **C5**, the 3′–OH substitution is hydrogen bonded to the backbone of Trp58. Moving this –OH substitution to 3′-position of B-ring, as in **C4**, leads to the breakage of the hydrogen bond with Trp58 and leaves the –OH at 3′-position of B-ring more exposed to the solvent, lowering the inhibitory activity of **C4**.

The absence of an –OH group at 3-position of C-ring, as in **C6**, and the presence of an –OH substitution in the 3′ position of B-ring, define a completely different mode of interaction. In **C6**, the lack of hydrogen bond donors in the C-ring precludes the binding within the catalytic triad. Instead, B-ring binds the catalytic triad with the two –OH substitutions hydrogen bonding Asp300, while Glu233 hydrogen bonds the 3′–OH (+1 subpocket). In this binding mode, the A- and C-rings occupy the site of the substrate cleft corresponding to the −1 and −2 subpockets of the leaving maltose, with the hydrogen of the –OH at 7-position of A-ring bonding Glu240.

In what concerns D group, the binding mode of **D11**, the flavonoid with the higher inhibitory activity, is almost identical to **B4**. The C-ring carbonyl group adjoined by the –OH group of A-ring allow the binding of A- and C-rings within the catalytic triad, establishing the same hydrogen bonding network with the active site residues Asp300, His299 and Glu233 (subpocket +1). This also means that B-ring of flavonoid **D11** binds in the same site as the B-ring of flavonoid **B4** and, like **B4**, His201 hydrogen bonds the –OH group at 3′-position of B-ring, while the 3′–OH hydrogen bonds Tyr151 (subpocket −1). On the other hand, the –OH group at 7-position of A-ring seems to establish no specific interactions, while the –Cl ion at 3-position of C-ring fills the cavity formed by Ile235 and Phe256 and hydrogen bonds the backbone of Ile234 [Figure 5(B)].

A similar binding mode is predicted when the –Cl ion at 3-position of C-ring of flavonoid **D11** is substituted by an -H or –OH group (**D7** and **D8**, respectively), but the inhibitory activity is decreased by two-fold and three-fold, respectively. The lower shape complementarity of **D7** and the resistance of binding the more electronegative oxygen in the pocket formed by Ile235 and Phe256 for flavonoid **D8** might explain the decreased inhibitory activity.

It is interesting to compare **D11** with **D14**, which bears an additional –Cl substitution at 8-position of A-ring. The binding mode of **D14** is exactly the same as **D11**, thus the 8–Cl substitution binds in the hydrophobic pocket formed by Leu162, Val163 and Leu165. Despite the increased shape complementarity, the IC_{50} value increased by twofold. Perhaps a similar sized but more hydrophobic group in this position would be advantageous.

Methoxylation of the B-ring in **D16**, **D17** and **D18** flipped the mode of binding, when compared to **D11**, and disrupted the interactions with the active site triad.

Among flavonoids of the E group, it was possible to observe that exchanging the sp^2 C2 carbon on the C-ring for a sp^3 carbon, has a dramatic effect on the inhibitory efficiency when compared to **D5**, **D7** and **D8**. The predicted mode of binding is the same for the three flavonoids, with the A- and C-rings occupying the +3 and +2 subpockets, while the B-ring occupies the +1 subpocket and hydrogen bonds Asp197. These three flavonoids clearly cannot occupy the subpockets +1 and −1 simultaneously to maximize the interaction with the catalytic triad, thus cannot inhibit α-amylase.

Discussion

The search for safer, specific and effective hypoglycaemic agents has been a concern of the scientific community. As a major player in the route leading to the assimilation of polysaccharides, the enzyme α-amylase plays an important role in controlling the postprandial blood glucose levels, thus being an interesting therapeutic target forT2DM. It was previously shown that the molecular models for the human and porcine pancreatic α-amylases are extremely similar[44], which enables the use of a porcine pancreatic α-amylase surrogate in microanalysis screening of novel α-amylase inhibitors [see Supporting Information, Figure 22(S), showing the superposition of the human and porcine enzymes].

In the present work, we tested the inhibition of pancreatic α-amylase activity by a panel of 40 flavonoids, that were divided among five groups (A–E) according to the respective parent structure scaffolds, with –OH, –OMe and/or –Cl substitutions, in order to elucidate a rationale for structure-activity relationship. The kinetic analysis, inhibitory mechanism determination and the corresponding molecular docking calculations were also applied, in order to explain the binding model of the most effective flavonoid of each group and the positive control, **acarbose**, regarding the three-dimensional structure of the α-amylase catalytic site. The obtained results clearly show that α-amylase inhibition by flavonoids varies with the nature, number and position of the substituents present in the derivative flavonoids' structure. From the studied flavonoids, the most effective compound found was the chlorinated flavone **D11**.

In what concerns the A group, the tested flavonoids (**A1**–**A6**) present low or no ability to inhibit α-amylase up to the highest tested concentration, 200 μM. Comparing the flavone **A1** with **A2**, **A3**, **A4** and **A5**, it is clear that addition of –OH groups at 3-position of C-ring and at 3′- and/or at 4′-positions of B-ring lead to no effect against the α-amylase activity. In addition, the substitution of –OH groups by –OMe groups at 3′-and 4′-positions of B-ring (**A6**) also did not lead to any effect for the inhibitory activity of α-amylase.

The most active flavonoid of B group was **B4** (IC_{50}= 148 ± 5 μM). Thus, the presence of a catechol group in B-ring, together with the –OH group at 5-position of A-ring, seems distinctive features for inhibitory activity, especially if a comparison is made with those derivatives without –OH groups in B-ring (**B1**) or with just one –OH group at 3′- or 4′-positions of B-ring (**B2** and **B3**, respectively). In addition, by comparing **A5** with **B4**, it can be noticed that the –OH group seems to be more favourable at 5-position of A-ring than at 3-position of C-ring for the intended activity. This is in accordance to previous studies showing that the hydroxylation of A- (at 5- and 7-positions) and B-rings (at 3′- and 4′-positions) of flavonoids improve their inhibitory effect against this enzyme[4,45–47].

The most effective flavonoid from the C group was **C5** (IC_{50}= 59 ± 4 μM), which was tested here for the first time. Taking into account the obtained result, it is possible to conclude that the presence of an –OH group at 3-position of C-ring, at 4′-position of B-ring and at 7-position of A-ring are advantageous for the inhibitory effect. Accordingly, it was previously reported, through the study of various structurally-related flavonoids, that the hydroxylation at 4′-position of B-ring and at 7-position of A-ring in flavonoids' scaffold is relevant for their inhibitory activity against α-amylase[4,21,45–48]. The importance of the presence of an –OH group at 4′-position in B-ring is also highlighted by comparing **C5** with **C4** (–OH group at 3′-position of the B-ring), since flavonoid **C4** presented low activity up to 200 μM, the highest tested concentration. The relevance of the presence of an –OH group at 3-position of C-ring together with an OH group at 4′-position of B-ring for the inhibition of α-amylase, is highlighted by the lack of inhibitory activity of **C3**, when compared with **C5**.

Concerning the studied methoxylated flavonoids, **C7** and **C8**, both presented no inhibitory activity against α-amylase, corroborating the idea that the inhibitory activity of flavonoids against α-amylase is not favoured through methoxyl substitution. Accordingly, it was previously described in *in vitro* studies that the methylation and methoxylation in the structure of flavonoids reduce the inhibitory activity towards α-amylase[45,46,48].

In D group we found the most active compound of this work, flavone **D11**, with a –Cl ion at 3-position of C-ring, a –OH group at 3′ and 4′-positions of B-ring and at 5- and 7-positions of A-ring, presenting an IC_{50} value = 44 ± 3 μM. Chlorinated flavonoids were previously reported by our research group as potent anti-inflammatory agents due to their ability to suppress mechanisms engaged at the onset and progression of inflammation[27,49]. Since both oxidative stress and inflammation play a key role in the pathogenesis of insulin resistance and dysfunction of β-cells[50], we selected these promising chlorinated flavonoids and studied their effectiveness as α-amylase inhibitors. In addition, these flavonoids already proved to be effective compounds in the inhibition of α-glucosidase[15] and PTP1B[16]. Comparing flavones **D11** and **D8** (quercetin) (IC_{50} = 138 ± 5 μM), it is possible to conclude that the substitution of the –Cl ion by an –OH group, at 3-position of C-ring, significantly decreased the α-amylase inhibition. Moreover, the removal of the –Cl ion from 3-position of C-ring also led to a decrease of the inhibitory activity of flavonoids, as it is possible to verify by comparing **D11** with **D7** (luteolin) (IC_{50} = 78 ± 3 μM). Furthermore, the addition of an –OH group at 3-position of C-ring [**D8** (quercetin)] significantly decreased the inhibitory activity towards α-amylase (138 ± 5 μM). The comparison of flavone **D11** with **D14** show that addition of a –Cl ion at 8-position of A-ring, led to a two-fold increase of the IC_{50} value (82 ± 4 μM). Thus, the presence of a –Cl ion at 3-position of the C-ring is especially advantageous for the ability of flavonoids to inhibit α-amylase activity.

Through the comparison of the results obtained for flavone **D9** (myricetin) (IC_{50} = 107 ± 6 μM) and **D8** (quercetin) (IC_{50} = 138 ± 5 μM), it was possible to observe that the addition of an –OH group at 5′-position of B-ring slightly enhanced α-amylase inhibition. The results of the present work are in accordance with those reported by Meng et al.[25] and Tadera et al.[21], who also studied the α-amylase inhibition by quercetin and myricetin and found that myricetin was slightly more effective than quercetin.

By comparing **D7** (luteolin) (IC_{50}= 78 ± 3 μM) and **D8** (quercetin) (IC50 = 138 ± 5 μM), it was also possible to conclude that the addition of an –OH group at 3-position of C-ring decreased the flavonoids' inhibitory activity towards α-amylase. The obtained results are in agreement with the study of Tadera et al.[21], who also tested the mentioned flavonoids. In the study of Kim et al.[23], the authors also observed that luteolin was an effective α-amylase inhibitor.

Comparing methoxylated flavonoids with their parent hydroxylated flavonoids, it is possible to conclude that the presence of

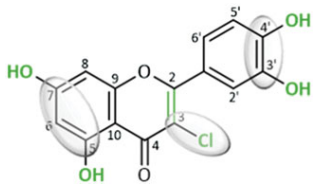


Figure 6. Potential substitution pattern of flavonoids contributing to the improvement of α-amylase inhibition.

–OMe groups decrease α-amylase inhibition, as it can be seen comparing flavone **D16** (chrysoeriol) and **D18** with **D7** (luteolin), and **D17** (acacetin) with **D5** (apigenin). These results are in agreement with previous studies found in literature[45–48,51].

A glycosylated flavonoid was also tested, **D19** (rutin), and the presence of a sugar in the flavonoid structure did not favour its ability to inhibit α-amylase activity. In this sense, the obtained results showed that the substitution of an –OH group [**D8** (quercetin)] by a rutinose [**D19** (rutin)] at 3-position of C-ring reduced the α-amylase inhibition by flavonoids. Accordingly, it was previously described in the literature that the glycosylation of flavonoids decreases their inhibitory effect against α-amylase, possibly due to the increasing molecular size and polarity, and to the nonplanar structure[45,46,48]. In addition, this conclusion is in agreement with the study of Li et al.[24], who tested the interaction between quercetin, rutin and isoquercetin with α-amylase, and found that the binding affinity of quercetin to the enzyme is higher when compared to rutin.

The flavanones from E group did not inhibit α-amylase, up to the highest tested concentration (200 μM). Comparing **E1** (naringenin) with **D5** (apigenin), **E2** (eriodictyol) with **D7** (luteolin) and **E3** (taxifolin) with **D8** (quercetin), it was observed that the lack of the C2 = C3 double bond did not favour the intended effect. This conclusion is in agreement with Tadera et al.,[21] who observed that naringenin was not able to inhibit porcine pancreatic α-amylase activity, at the highest tested concentration of 0.5 mM. It is well known that the C2 = C3 double bond allows an easier entry of the molecules into the hydrophobic pockets of α-amylase since the presence of this double bond favours the near-planarity of flavonoids[4,45,46,48].

Acarbose is a pseudotetrasaccharide produced by *Actinoplanes* sp. fermentation and its structure comprises a hydroxymethyl conduritol residue α-(1,4) linked to a 4-amino-4,6-dideoxyglucose which is, in turn, α-(1,4) linked to maltose[52,53]. It is the most widely prescribed α-glucosidase and α-amylase inhibitor. However, it was previously shown that only a mild inhibition of pancreatic α-amylase is recommended in order to avoid gastrointestinal side-effects as a result of excessive bacterial fermentation of carbohydrates in colon[11], that has been shown as a limiting factor for T2DM treatment, in some countries[11]. In the present work, the IC_{50} values found for the tested flavonoids were higher than the IC_{50} value found for acarbose, which is 1.3 ± 0.2 μM. The mild pancreatic α-amylase inhibition prevents the abnormal bacterial fermentation of carbohydrates in the colon and consequent gastrointestinal adverse effects, such as abdominal distention, flatulence and diarrhea[11]. Moreover, it was previously shown by our research group that, in general, the tested flavonoids were more effective concerning the inhibition of α-glucosidase than the inhibition of α-amylase[15]. As such, the obtained results suggest that flavonoids could be promising α-amylase inhibitors and cause less gastrointestinal side-effects. Interestingly, it was found that the combination of baicalein or apigenin with acarbose showed additive inhibition of α-amylase at lower concentrations and antagonistic inhibition at higher concentrations. On the other hand, the combination of baicalein with acarbose synergistically inhibits α-glucosidase and lowers HGPP[54].

After the study of α-amylase inhibition by flavonoids, we selected the most active flavonoid of each group (**B4**, **C5** and **D11**) and the positive control, **acarbose**, to test their type of inhibition.

It was described that the discrimination between mechanisms of enzymatic inhibition based on linear transformation of Michaelis–Menten equation is less accurate than the use of the nonlinear regression, since the transformation can also mask the variability inherent to kinetic data[29,30]. By the same reason, modelling on the means of repeated experiments instead of using individual data can also lead to misinterpreted mechanisms. In that sense, for the enzyme kinetic analysis, it was used the nonlinear regression to determine kinetic parameters and the type of inhibition of the most active flavonoids of each group. For this purpose, it was applied the Solver supplement of Microsoft Office Excel™. By means of iterative guesses for variables in each model equation, Solver was asked to minimize the sum of squared residuals between individual raw experimental results and the corresponding values generated by the model.

In accordance, kinetic analysis consisted in the sequential fitting of available models (without inhibition, competitive inhibition, noncompetitive inhibition, uncompetitive inhibition and mixed inhibition), and in the estimation of the kinetic constants (parameters) values. It was possible to observe by the analysis of the sum of the square errors values from the different models, and by the application of F-test and AIC test, that **acarbose** presents a mixed-type inhibition [see Supporting Information, Figures 1(S)–5(S)], and flavonoids **B4** [see Supporting Information, Figures 6(S)–10(S)], **C5** (see Supporting Information, Figures 11(S)–15(S)) and **D11** [see Supporting Information, Figures 16(S)–20(S)] present a competitive inhibition. Accordingly, Kim et al.[55] and Yoon and Robyt,[52] also described acarbose as a mixed type inhibitor of α-amylase. To the best of our knowledge, the inhibition type of flavonoids **B4**, **C5** and **D11** were studied here for the first time.

Once defined the best kinetic model for **acarbose** and for flavonoids **B4**, **C5** and **D11**, the kinetic constants were estimated: V_{max}, K_m, K_{ic} and K_{iu}. The calculation of the Ki value is considered a useful tool to compare the potency of inhibitors. The K_{ic} values of the selected compounds demonstrated the following order: **B4**>**C5**>**D11**>**acarbose**. This means that, as expected, **acarbose** has the highest binding affinity to α-amylase, followed by flavonoids **D11**, **C5** and **B4**.

The prediction of the binding mode of the most active flavonoids towards α-amylase in each group showed that one of the most important aspects shared by them is a motif bearing a carbonyl group next to a –OH group. This allows the carbonyl group to accept a hydrogen bond from Glu233 and the –OH group to donate a hydrogen bond to Asp300 in the region between the –1 subpocket and +1 subpocket. The interactions with these groups can be further enhanced with a hydrogen bond between His299 and the -OH group.

The relative positioning of the -OH and carbonyl groups determine, to a great extent, the orientation of the flavonoid. Thus, the flavonoid part on the side of the carbonyl will be preferably oriented towards the +1 subpocket, while the part on the -OH side will be preferably oriented towards the –1 subpocket.

Hydrophobic groups seem to be more tolerated on the –1/–2 subpockets than on the +1/+2 subpockets, where they can interact with Tyr62, Leu162, Val163 and Leu165. Hydrophobic groups on the +1/+2 subpockets, on the other hand, can establish hydrogen bonds with Tyr151 and His201.

Halogen substitutions or hydrophobic groups might be used next to the carbonyl to explore the interaction with Ile235 and Phe256, as the best flavonoid tested herein shows.

Conclusion

In the present study, the inhibition of pancreatic α-amylase by a panel of 40 flavonoids (A–E groups) was tested, in order to establish an accurate structure-activity relationship. For that purpose, the kinetic analysis, the inhibitory mechanism determination and the corresponding molecular docking calculations were applied, to explain the binding model of the most effective flavonoid of each group to the three-dimensional structure of α-amylase catalytic site. Thus, from the obtained results, it may be concluded that α-amylase inhibition by flavonoids is strongly dependent on the nature and position of the substituents. The most effective flavonoid was the chlorinated flavone **D11**. As such, the presence of an –OH group at 5- and 7-positions of A-ring and in the 3′- and 4′-positions of B-ring, and the presence of a –Cl ion at 3-position of C-ring, as well as the C2 = C3 double bond, have shown to be favourable for the intended effect (Figure 6). This flavonoid presented a competitive type of inhibition. The promising α-amylase inhibitory activity of some of the studied flavonoids should be deeply explored allowing these compounds to be considered as possible alternatives for the management of PPHG and consequently T2DM.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was financed by FEDER–Fundo Europeu de Desenvolvimento Regional through the COMPETE 2020–Operacional Programme for Competitiveness and Internationalisation (POCI), and by Portuguese funds through FCT–Fundação para a Ciência e a Tecnologia in the framework of the project POCI-01–0145-FEDER-029241. Carina Proença acknowledges FCT the financial support for the PhD grant [SFRH/BD/116005/2016], in the ambit of "QREN–POPH–Tipologia 4.1–FormaçãoAvançada", co-sponsored by Fundo Social Europeu (FSE) and by national funds of Ministério da Ciência, Tecnologia e Ensino Superior (MCTES).

ORCID

Carina Proença https://orcid.org/0000-0003-0859-6526
Marisa Freitas https://orcid.org/0000-0001-9114-9967
Daniela Ribeiro https://orcid.org/0000-0001-9733-7934
Sara M. Tomé http://orcid.org/0000-0002-9021-2909
Eduardo F. T. Oliveira https://orcid.org/0000-0003-4637-2099
Matilde F. Viegas https://orcid.org/0000-0002-4472-9691
Alberto N. Araújo http://orcid.org/0000-0001-5348-4924
Maria J. Ramos http://orcid.org/0000-0002-7554-8324
Artur M. S. Silva http://orcid.org/0000-0003-2861-8286
Pedro A. Fernandes http://orcid.org/0000-0003-2748-4722
Eduarda Fernandes http://orcid.org/0000-0001-6424-0976

References

1. Butterworth PJ, Warren FJ, Ellis PR. Human α-amylase and starch digestion: an interesting marriage. Starch–Stärke 2011;63:395–405.
2. Patel H, Royall PG, Gaisford S, et al. Structural and enzyme kinetic studies of retrograded starch: inhibition of α-amylase and consequences for intestinal digestion of starch. Carbohydr Polym 2017;164:154–61.
3. Brayer GD, Sidhu G, Maurus R, et al. Subsite mapping of the human pancreatic α-amylase active site through structural, kinetic, and mutagenesis techniques. Biochemistry 2000;39:4778–91.
4. Lo Piparo E, Scheib H, Frei N, et al. Flavonoids for controlling starch digestion: structural requirements for inhibiting human α-amylase. J Med Chem 2008;51:3555–61.
5. Sales PM, Souza PM, Simeoni LA, et al. A-amylase inhibitors: a review of raw material and isolated compounds from plant source. J Pharm Pharm Sci 2012;15:141–83.
6. Liu D, Gao H, Tang W, Nie S. Plant non-starch polysaccharides that inhibit key enzymes linked to type 2 diabetes mellitus. Ann N Y Acad Sci 2017;1401:28–36.
7. Martinez-Gonzalez AI, Diaz-Sanchez AG, Rosa LA, et al. Polyphenolic compounds and digestive enzymes: *in vitro* non-covalent interactions. Molecules 2017;22:1–24.
8. Janecek S, Svensson B, MacGregor EA. α-Amylase: an enzyme specificity found in various families of glycoside hydrolases. Cell Mol Life Sci 2014;71:1149–70.
9. International Diabetes Federation. IDF Diabetes Atlas. In IDF Diabetes Atlas. 8th ed. 2017. Available from: http://diabetesatlas.org/IDF_Diabetes_Atlas_8e_interactive_EN/
10. American Diabetes Association. Classification and diagnosis of diabetes: standards of medical care in diabetes–2018. Diabetes Care 2018;41:S13–S27.
11. Etxeberria U, de la Garza AL, Campion J, et al. Antidiabetic effects of natural plant extracts via inhibition of carbohydrate hydrolysis enzymes with emphasis on pancreatic alpha amylase. Expert Opin Ther Targets 2012;16:269–97.
12. Kahn SE, Cooper ME, Del Prato S. Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. Lancet 2014;383:1068–83.
13. Wedick NM, Pan A, Cassidy A, et al. Dietary flavonoid intakes and risk of type 2 diabetes in US men and women. Am J Clin Nutr 2012;95:925–33.
14. Liao X, Wang X, Li H, et al. Sodium-glucose cotransporter 2 (SGLT2) inhibitor increases circulating zinc-α2-glycoprotein levels in patients with type 2 diabetes. Sci Rep 2016;6:1–12.
15. Proenca C, Freitas M, Ribeiro D, et al. α-Glucosidase inhibition by flavonoids: an *in vitro* and *in silico* structure–activity relationship study. J Enzyme Inhib Med Chem 2017;32:1216–28.
16. Proenca C, Freitas M, Ribeiro D, et al. Inhibition of protein tyrosine phosphatase 1B by flavonoids: a structure–activity relationship study. Food Chem Toxicol 2018;111:474–81.
17. Nicolle E, Souard F, Faure P, Boumendjel A. Flavonoids as promising lead compounds in type 2 diabetes mellitus: molecules of interest and structure–activity relationship. Curr Med Chem 2011;18:2661–72.
18. Vinayagam R, Xu B. Antidiabetic properties of dietary flavonoids: a cellular mechanism review. Nutr Metab 2015;12:1–20.
19. Xiao JB, Hogger P. Dietary polyphenols and type 2 diabetes: current insights and future perspectives. Curr Med Chem 2015;22:23–38.
20. Kawser HM, Abdal DA, Han J, et al. Molecular mechanisms of the anti-obesity and anti-diabetic properties of flavonoids. Int J Mol Sci 2016;17:569–32.

21. Tadera K, Minami Y, Takamatsu K, Matsuoka T. Inhibition of alpha-glucosidase and alpha-amylase by flavonoids. J Nutr Sci Vitaminol 2006;52:149–53.
22. Gu C, Zhang H, Putri CY, Ng K. Evaluation of α-amylase and α-glucosidase inhibitory activity of flavonoids. Int J Food Nutr Sci 2015;2:1–6.
23. Kim JS, Kwon CS, Son KH. Inhibition of alpha-glucosidase and amylase by luteolin, a flavonoid. Biosci Biotechnol Biochem 2000;64:2458–61.
24. Li Y, Gao F, Gao F, et al. Study on the interaction between 3 flavonoid compounds and α-amylase by fluorescence spectroscopy and enzymatic kinetics. J Food Sci 2009;74: C199–203.
25. Meng Y, Su A, Yuan S, et al. Evaluation of total flavonoids, myricetin, and quercetin from *Hovenia dulcis* thunb. as inhibitors of α-amylase and α-glucosidase. Plant Foods Hum Nutr 2016;71:444–9.
26. Ribeiro D, Freitas M, Tome SM, et al. Modulation of human neutrophils' oxidative burst by flavonoids. Eur J Med Chem 2013;67:280–92.
27. Freitas M, Ribeiro D, Tome SM, et al. Synthesis of chlorinated flavonoids with anti-inflammatory and pro-apoptotic activities in human neutrophils. Eur J Med Chem 2014;86:153–64.
28. Trinh BTD, Staerk D, Jäger AK. Screening for potential α-glucosidase and α-amylase inhibitory constituents from selected Vietnamese plants used to treat type 2 diabetes. J Ethnopharmacol 2016;186:189–95.
29. Bezerra RMF, Fraga I, Dias AA. Utilization of integrated Michaelis–Menten equations for enzyme inhibition diagnosis and determination of kinetic constants using Solver supplement of Microsoft Office Excel. Comput Methods Programs Biomed 2013;109:26–31.
30. Dias AA, Pinto PA, Fraga I, Bezerra RMF. Diagnosis of enzyme inhibition using Excel Solver: a combined dry and wet laboratory exercise. J Chem Educ 2014;91:1017–21.
31. Motulsky H, Christopoulos A, Fitting models to biological data using linear and nonlinear regression: a pratical guide to curve fitting. New York: Oxford University Press; 2004.
32. Burguillo FJ, Wright AJ, Bardsley WG. Use of the F test for determining the degree of enzyme-kinetic and ligand-binding data. A Monte Carlo simulation study. Biochem J 1983; 211:23–34.
33. Yamaoka K, Nakagawa T, Uno T. Application of Akaike's information criterion (AIC) in the evaluation of linear pharmacokinetic equations. J Pharmacokinet Biopharm 1978;6: 165–75.
34. Naes T, Isakson T, Fearn T, Davies T, A user-friendly guide to multivariate calibration and classification. Chichester: NIR Publications; 2002.
35. Qian M, Nahoum V, Bonicel J, et al. Enzyme-catalyzed condensation reaction in a mammalian alpha-amylase. High-resolution structural analysis of an enzyme-inhibitor complex. Biochemistry 2001;40:7700–9.
36. Bietz S, Urbaczek S, Schulz B, Rarey M. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. J Cheminform 2014; 6:1–12.
37. Case DA, Darden TA, Cheatham TE, et al. Amber 12. In. Amber 12. San Francisco: University of California; 2012.
38. Hornak V, Abel R, Okur A, et al. Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins 2006;65:712–25.
39. Kirschner KN, Yongye AB, Tschampel SM, et al. Glycam06: a generalizable biomolecular force field. J Comput Chem 2008;29:622–55.
40. Trott O, Olson AJ. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 2010;31: 455–61.
41. O'Boyle NM, Morley C, Hutchison GR. Pybel: a python wrapper for the openbabel cheminformatics toolkit. Chem Cent J 2008;2:5–7.
42. Gasteiger J, Marsili M. Iterative partial equalization of orbital electronegativity–a rapid access to atomic charges. Tetrahedron 1980;36:3219–28.
43. Jones G, Willett P, Glen RC, et al. Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267:727–48.
44. Qian M, Spinelli S, Driguez H, Payan F. Structure of a pancreatic α-amylase bound to a substrate analogue at 2.03 a resolution. Protein Sci 2008;6:2285–96.
45. Cao H, Chen X. Structures required of flavonoids for inhibiting digestive enzymes. Anticancer Agents Med Chem 2012; 12:929–39.
46. Xiao J, Ni X, Kai G, Chen X. A review on structure–activity relationship of dietary polyphenols inhibiting α-amylase. Crit Rev Food Sci Nutr 2013;53:497–506.
47. Gonzales GB, Smagghe G, Grootaert C, et al. Flavonoid interactions during digestion, absorption, distribution and metabolism: a sequential structure–activity/property relationship-based approach in the study of bioavailability and bioactivity. Drug Metab Rev 2015;47:175–90.
48. Xu W, Shao R, Xiao J. Is there consistency between the binding affinity and inhibitory potential of natural polyphenols as α-amylase inhibitors? Crit Rev Food Sci Nutr 2016;56: 1630–9.
49. Proença C, Ribeiro D, Soares T, et al. Chlorinated flavonoids modulate the inflammatory process in human blood. Inflammation 2017;40:1155–65.
50. Pitocco D, Tesauro M, Alessandro R, et al. Oxidative stress in diabetes: implications for vascular and other complications. Int J Mol Sci 2013;14:21525–50.
51. Asghari B, Salehi P, Sonboli A, Nejad Ebrahimi S. Flavonoids from *salvia chloroleuca* with α-amylase and α-glucosidase inhibitory effect. Iran J Pharm Res 2015;14:609–15.
52. Yoon S-H, Robyt JF. Study of the inhibition of four alpha amylases by acarbose and its 4iv-α-maltohexaosyl and 4iv-α-maltododecaosyl analogues. Carbohydr Res 2003;338:1969–80.
53. Rosak C, Mertes G. Critical evaluation of the role of acarbose in the treatment of diabetes: patient considerations. Diabetes Metab Syndr Obes 2012;5:357–67.
54. Zhang BW, Li X, Sun WL, et al. Dietary flavonoids and acarbose synergistically inhibit α-glucosidase and lower postprandial blood glucose. J Agric Food Chem 2017;65: 8319–30.
55. Kim M, Lee S, Lee H, et al. Comparative study of the inhibition of α-glucosidase, α-amylase, and cyclomaltodextrin glucanosyltransferase by acarbose, isoacarbose, and acarviosine–glucose. Arch Biochem Biophys 1999;371: 277–83.

8.2. Animal Fatty Acid Synthase: a Chemical Nanofactory

Rui P. P. Neves[a], Pedro Ferreira[a], Fabiola E. Medina [b], Pedro Paiva[a], João P. M. Sousa[a], Matilde F. Viegas[a], Maria J. Ramos[a], and Pedro A. Fernandes[a]

[a] LAQV@REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

[b] Departamento de Ciencias Químicas, Facultad de Ciencias Exactas, Universidad Andres Bello, Autopista Concepción-Talcahuano, 7100 Talcahuano, Chile

---

Paiva, P., Medina, F. E., **Viegas, M. F.**, Ferreira, P., Neves, R. P. P., Sousa, J. P. M., Ramos, M. J., & Fernandes, P. A. (2021). ***Animal fatty acid synthase: A chemical nanofactory***. In Chemical Reviews, 121(15), 9502–9553. https://doi.org/10.1021/acs.chemrev.1c00147

CHEMICAL REVIEWS
pubs.acs.org/CR Review

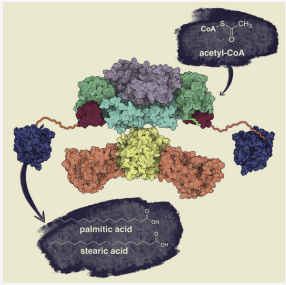
Animal Fatty Acid Synthase: A Chemical Nanofactory

Pedro Paiva,[⊥] Fabiola E. Medina,[⊥] Matilde Viegas,[⊥] Pedro Ferreira,[⊥] Rui P. P. Neves,[⊥] João P. M. Sousa,[⊥] Maria J. Ramos, and Pedro A. Fernandes*

Cite This: https://doi.org/10.1021/acs.chemrev.1c00147 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Fatty acids are crucial molecules for most living beings, very well spread and conserved across species. These molecules play a role in energy storage, cell membrane architecture, and cell signaling, the latter through their derivative metabolites. *De novo* synthesis of fatty acids is a complex chemical process that can be achieved either by a metabolic pathway built by a sequence of individual enzymes, such as in most bacteria, or by a single, large multi-enzyme, which incorporates all the chemical capabilities of the metabolic pathway, such as in animals and fungi, and in some bacteria. Here we focus on the multi-enzymes, specifically in the animal fatty acid synthase (FAS). We start by providing a historical overview of this vast field of research. We follow by describing the extraordinary architecture of animal FAS, a homodimeric multi-enzyme with seven different active sites per dimer, including a carrier protein that carries the intermediates from one active site to the next. We then delve into this multi-enzyme's detailed chemistry and critically discuss the current knowledge on the chemical mechanism of each of the steps necessary to synthesize a single fatty acid molecule with atomic detail. In line with this, we discuss the potential and achieved FAS applications in biotechnology, as biosynthetic machines, and compare them with their homologous polyketide synthases, which are also finding wide applications in the same field. Finally, we discuss some open questions on the architecture of FAS, such as their peculiar substrate-shuttling arm, and describe possible reasons for the emergence of large megasynthases during evolution, questions that have fascinated biochemists from long ago but are still far from answered and understood.



CONTENTS

1. Historical Perspective B
1.1. Fatty Acid Biosynthesis B
1.2. Fatty Acid Synthase, Cancer, and Obesity C
2. Animal Fatty Acid Synthase: Architecture and Function D
2.1. Architecture of Animal FAS D
2.1.1. Expression, Folding, and Assembly Elasticity D
2.1.2. Overall Architecture and Catalysis D
2.1.3. Head-to-Tail vs Head-to-Head Orientation E
2.1.4. Structural Scaffolding and Flexibility E
2.2. Physiological Role of Animal FAS and Its Role in Therapeutics F
2.2.1. Mechanisms of Animal FAS Expression and Implications in Disease F
2.2.2. Drug Design Targeting Animal FAS G
3. The Catalytic Mechanism of Animal FAS H
3.1. Reaction Cycle of the MAT Domain H
3.1.1. Substrate Binding and Promiscuity H
3.1.2. Active-Site Structure and Catalytic Mechanism I
3.2. Reaction Cycle of the KS Domain K
3.2.1. Structure and Function of the Animal KS Domain K
3.2.2. Catalytic Machinery of the KS Domain L
3.2.3. Role of the Animal KS Domain in the Fatty Acid Chain-Length Regulation P
3.3. Reaction Cycle of the KR Domain Q
3.3.1. Biochemical Characterization of Animal KR Q
3.3.2. Active-Site Structure and Catalytic Mechanism of KR Q
3.4. Reaction Cycle of the DH Domain S
3.4.1. Biochemical and Structural Characterization S
3.4.2. Active-Site Structure and Catalytic Mechanism S
3.5. Reaction Cycle of the ER Domain T
3.5.1. Biochemical and Structural Characterization T
3.5.2. Active-Site Structure and Catalytic Mechanism T
3.6. Reaction Cycle of the TE Domain W

Received: February 19, 2021

ACS Publications © XXXX American Chemical Society

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

3.6.1. The TE Superfamily W
3.6.2. Animal FAS TE1 and TE2 W
3.6.3. Structure Overview W
3.6.4. Catalytic Mechanism X
3.6.5. Substrate Binding Sites and Substrate Specificity Y
3.6.6. Interplay between TE1 and TE2 Z
3.7. Overall Reaction Cycle of FAS Z
3.7.1. Kinetic Rates of the Individual Reactions of Fatty Acid Synthesis Z
3.7.2. Molecular View of the Multi-enzyme Reaction Mechanism AA
3.7.3. Chemical Highlights of the Reaction Mechanism of FAS AA
4. Why Has Evolution Allowed the Emergence of Complex Megasynthases? AB
4.1. Animal, Fungal, and Bacterial FAS AB
4.2. Role of the ACP in Enzyme Efficiency AC
5. Relation of FAS with PKSs and the Synthesis of Secondary Metabolites AD
5.1. PKSs AD
5.2. Evolutionary Relationship between FASs and PKSs AE
5.3. Polyketide Synthesis and the PKS Machinery (Megasynthases vs Monofunctional Synthases) AE
5.4. Short Reference to the Engineering of PKSs AF
6. Biotechnological Applications of FASs AG
6.1. Metabolic Engineering of FASs AG
6.1.1. Metabolic Engineering in Bacteria AG
6.1.2. Metabolic Engineering in Yeast AH
6.1.3. Orthogonal Fatty Acid Biosynthesis AH
6.2. Enzyme Engineering of FASs AH
6.2.1. FAS Engineering for Biofuel Production AH
6.2.2. FAS Engineering for Polyketide Synthesis AI
7. Conclusion AJ
Associated Content AK
Supporting Information AK
Author Information AK
Corresponding Author AK
Authors AK
Author Contributions AK
Notes AK
Biographies AK
Acknowledgments AL
References AL

1. HISTORICAL PERSPECTIVE

1.1. Fatty Acid Biosynthesis

Fatty acid biosynthesis is an anabolic pathway responsible for the manufacture of fatty acids—a crucial process for most living cells, highly conserved across species. Fatty acids are versatile molecules involved in energy storage[1] membrane composition,[2] and their derivatives possess signaling properties.[3,4]

The initial discoveries on fat biosynthesis started in the middle of the 19th century. One of the first experiments on this matter, by Rittenberg and Bloch, demonstrated that deuterium-labeled C_2 units were used for lipid biosynthesis.[5] About a decade following this initial landmark, Klein and Lippman observed a correlation between the amount of acetyl-CoA and lipid synthesis, in both yeast and animal cells, demonstrating the involvement of acetyl-CoA in lipid biosynthesis.[6,7]

Back then, fatty acid synthesis was thought to happen in the mitochondria, carried out by the same enzymes responsible for the β-oxidation of fatty acids. Wakil later found out that this observation was only true for the longer-chain fatty acids, and that *de novo* fatty acid synthesis occurred in the soluble fraction of cells by the intermediate of a malonic acid derivative.[8] Meanwhile, also in 1958, Brady discovered that malonyl-CoA is an exclusive intermediate for the formation of C_{16} palmitate, thus establishing the non-mitochondrial system as the main pathway for fatty acid synthesis. Nonetheless, Zhang et al. later reported a distinct system for fatty acid synthesis that can also be found in the mitochondria of animal cells, which employs enzymes that resemble those that participate in typical type II FAS systems.[9,10]

Interestingly enough, Norman Radin, who was not directly linked to the fatty acid biosynthesis research, made an important contribution when he intervened in Salih Wakil's talk at the 1958 Gordon Conference on Lipid Metabolism held in New Hampshire, suggesting that the CO_2 stimulation could mean that malonyl-CoA was involved in the biosynthetic route. Shortly after this event, the cytosolic malonyl-CoA pathway was regarded as the major pathway for fatty acid biosynthesis,[8,10] proving that sharing knowledge between different research areas can lead to major scientific breakthroughs.

Wakil and Gibson also determined evidence for this cytosolic system for fatty acid synthesis using an avian liver fraction, free of enzymes involved in the β-oxidation pathway. They observed that this fraction could still perform the biosynthesis of fatty acids when provided with acetyl-CoA, ATP, Mn^{2+}, CO_2, and NADPH.[8,11−14]

The global equation for fatty acid synthesis was then established in 1964 by Wakil and co-workers[15] and described in eq 1. In this condensation reaction, acetyl-CoA acts as primer for the reaction, corresponding to carbons 15 and 16 of the palmitate molecule, while malonyl-CoA, formed by acetyl-CoA carboxylase, is responsible for the remaining 14 carbons.

$$\text{acetyl-CoA} + 7\text{malonyl-CoA} + 14\text{NADP}^+ + 14\text{H}^+ \rightleftharpoons \text{palmitic acid} + 7\text{CO}_2 + 8\text{CoA} + 14\text{NADP} + 6\text{H}_2\text{O} \quad (1)$$

Simultaneously, the molecular details of fatty acid biosynthesis were being unveiled by the Vagelos group, using *Escherichia coli* as the model organism for their studies. They found out that the acyl intermediates of this pathway were bound to a protein that contains a 4′-phosphopantetheine (PNS) prosthetic group, the acyl carrier protein (ACP). This protein is responsible for carrying these intermediates from one to another of the enzymes of the *E. coli* fatty acid biosynthetic system.[16,17] Other than the ACP, the remaining enzymes of the *E. coli* FAS metabolic pathway were characterized in the 1960s. The Vagelos group was again responsible for the identification of the catalytic components with malonyl/acetyl transferase (MAT), β-ketoacyl synthase (KS), β-ketoacyl reductase (KR), and dehydratase (DH)[18,19] activities, and Wakil and co-workers identified the enoyl reductase (ER) activity.[20] They discovered that each of these catalytic components were standalone enzymes, each capable of catalyzing their respective step in the biosynthesis of fatty acids.

The ACP was identified later also in rat liver[21] and *Saccharomyces cerevisiae*[22] FAS biosynthetic systems, as well as

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

the respective catalytic components.[23] Unlike the prokaryotic FAS, the ACP and the other catalytic components of pigeon liver FAS were found to be enzymatic domains assembled in a large multi-enzyme complex, which could be dissociated into two polypeptides.[24]

Based on the *in vitro* behavior of both types of FAS systems, Bloch and co-workers formulated a classification that holds up until today: the type I FAS system, essentially found in higher eukaryotes, fungi, *Corynebacterium*, *Mycobacterium*, and *Nocardia* (CMN) group bacteria, where the individual catalytic units have merged to form a single multi-functional polypeptide FAS, possessing all catalytic activities and expressed from a single gene,[25] and the type II FAS system, seen in plants, bacteria, and mitochondria,[26] characterized by its intervening enzymes existing as discrete, free-standing, individual entities that form a metabolic pathway for fatty acid synthesis.

Even within the type I FAS systems, different assemblies of the system can be found, and the comparison between the animal and the fungal FAS is a good example of such variability. While the animal FAS was found to be a homodimer with a molecular weight of about 0.5 MDa, the fungal FAS is a dodecamer, with 6 pairs of non-identical subunits and 2.6 MDa of molecular weight.[25]

In the late 1960s, the ACP from *E. coli* became the first protein of the type II FAS system to be sequenced.[27,28] The primary structures of the enzymes KS I and DH of *E. coli* FAS system were obtained in the 1980s,[29,30] and the remainder of the enzymes of the *E. coli* type II FAS system were sequenced in the following decade,[31] alongside their three-dimensional structures, which were all resolved by X-ray crystallography except for the ACP structure, which was obtained by nuclear magnetic resonance.[32−40]

The elucidation of the primary structures of FAS from yeast, rat, and chicken was concluded by the late 1980s.[41−45] Despite both fungal FAS and animal FAS being FAS type I systems, the publication of their sequences and genetic studies showed that they evolved along two separate lines and have considerably different architectures.[41,43,44,46]

The complexity of the multi-enzyme architecture of the type I FAS system makes the determination of its 3D structure a critical point for the understanding of its way of action. This motivated structural studies to be conducted alongside the biochemical and genetic ones since the 1960s.[47] The first studies conducted to analyze the structure of type I FAS (from yeast) were based on images obtained by electron microscopy (EM)[48] and small-angle X-ray scattering (SAXS) in the 1960s and early 1970s.[49] Those images revealed particles with a hollow interior surrounded by a ring with 210 × 250 Å.[48,49] In 1978, the first model of the architecture of the type I fungal FAS of yeast was proposed.[50] In the following years, other proposed models corroborated the general structure of fungal FAS[51−53] of yeast and, in the late 1990s, it was concluded that the FAS multi-enzyme has an internal D3 symmetry, using 3D cryo-EM.[54,55] The unavailability of high-resolution X-ray structures impeded identifying the individual domains within the 3D architecture of the whole multi-enzyme.[54] The first X-ray structures of the fungal FAS from the fungus *Thermomyces lanuginosus* were reported in 2006,[56] with a much better resolution (5 Å) than the previously determined structures, but still far from the desired high resolution. In the following two years Lomakin and colleagues[57] and Leibundgut and co-workers[58] obtained crystals that diffracted at 4 and 3 Å resolution, respectively.

Similarly, for animal FAS, the earliest structure predictions, in the 1980s, were based on EM images and small-angle neutron scattering (SANS).[59−61] Those images supported the old head-to-tail model, which stated that the two polypeptide chains of animal FAS lay side by side, in an antiparallel orientation. When genetic and biophysical approaches challenged the head-to-tail model and suggested a head-to-head instead, mainly due to the observation that the KS domains of the two polypeptide chains are in close proximity, which is incompatible with a side-by-side antiparallel orientation,[41,62−64] the need for higher resolution structures was made evident (for a detailed description of the head-to-tail and the head-to-head model see section 2 of this Review). Unfortunately, the high conformational variability and flexibility of animal FAS seriously limit the 3D reconstruction of the protein to resolutions around 30 Å obtained with electron cryomicroscopy, insufficient for recognizing the individual domains.[59,65] Asturias and Smith deduced that if they used mutant FAS in the presence of substrates, they could limit the conformational variability of the enzyme and obtain better images, which allowed them to further corroborate the head-to-head model.[66] Unfortunately, despite the smaller size of animal FAS comparatively to fungal FAS, the high number of possible conformations, frequently without a 2-fold symmetry, made it a more difficult protein for crystallization.[60] Even though crystals were obtained as early as the late 1960s, when Oesterhelt et al. obtained crystals of yeast FAS,[67] they remained unsuitable for high-resolution X-ray crystallographic analysis until the 21st century. At last, Maier et al., in 2006, succeeded in obtaining crystals that diffracted at 4.5 Å resolution.[68] As the structure of each domain of the type II FAS system was previously solved with much higher resolution than the one obtained for the full type I FAS multi-enzyme, the structure of the latter was obtained with the help of the individual type II FAS enzymes homologues that served as templates for each type I FAS catalytic domain. The final structure of porcine type I FAS displayed the KS, MAT, DH, KR, and ER catalytic domains Only the KR domain was shown to be different from its type II FAS homologue enzyme, which is a tetramer. The other domains turned out to be similar to their type II FAS system counterparts. Due to the high mobility of ACP and TE, the authors were unable to find appropriate electron density to obtain a crystal structure of those domains. However, the structures of both individual ACP and TE domains had been previously proposed, in 2003 and 2004 respectively, using NMR in the case of ACP,[69] and X-ray crystallography in the case of TE.[70]

1.2. Fatty Acid Synthase, Cancer, and Obesity

The role that FAS has in the growth of malignant tumors is one of the most, if not the most, important aspect motivating the research of the properties and way of action of this complex protein. The association of FAS with cancer dates back to the 1950s when a high level of fatty acid synthesis in cancer cells was observed.[71,72] It took 30 years to demonstrate that tumor fatty acid synthesis comes almost exclusively from *de novo* synthesis.[73] In the 1990s, it was observed that breast cancer cells exhibit abnormally high levels of FAS expression and activity.[74] Increased levels of FAS were subsequently observed in prostate,[75,76] colon,[77] ovary,[78] and endometrium[79] aggressive tumors. FAS was also identified as a potential marker for the diagnosis or prognosis of cancer due to being detected in the blood of cancer patients.[80] The discovery that FAS plays a crucial role in tumor development, due to the observation that inhibition of the enzyme or its genetic knockdown leads to the

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

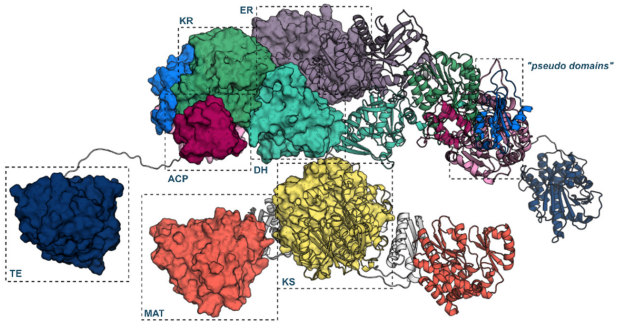


Figure 1. Representation of the animal FAS homodimer head-to-head model, as obtained from homology modeling.[118] Each monomer pair is represented in surface and cartoon. ACP corresponds to the mobile acyl-carrier protein domain carrying the substrate to the catalytic malonyl/acetyl transferase (MAT), β-ketoacyl synthase (KS), β-ketoacyl reductase (KR), dehydratase (DH), enoyl reductase (ER), and finally thioesterase (TE) domains. The ACP domain is represented bound to KR. It binds all the other domains during the multi-enzyme cycle. The TE domain is highly mobile, and the position illustrated is merely illustrative. The non-catalytic methyltransferase- and β-ketoreductase-like "pseudo-domains" are also highlighted.

(selective) apoptosis of cancer cells,[81−85] made it a very promising therapeutic target. The potential of FAS as a therapeutic target for aggressive tumors led to the emergence of many patents for the use of FAS inhibitors in the treatment of various cancers.[86−97]

In the early 2000s, FAS also started to be seen as a central point for the development of obesity.[98,99] Increased levels of FAS in adipose tissue appear to cause an excess of energy intake and fat accumulation.[100] Consequently, some of the FAS inhibitors initially developed for the treatment of cancer are also anti-obesity agents, and vice versa.[101−103]

In summary, the research on FAS biosynthetic systems has over seven decades of history. Many progresses have been made in the elucidation of their structures, their physiological role, and their reaction mechanisms. The role of FAS in cancer is now starting to be better understood. However, the understanding of the FAS overall mechanism at the atomic level is far from established. Accordingly, many questions remain about the mechanism of each FAS enzyme/domain, and the elaborate process of substrate transport across domains in type I FAS systems. The evolutive advantages of the complex architecture of FAS are still merely hypothesized, and efficient *in vivo* inhibition of FAS for therapeutic purposes is still to be achieved. The research community has come a long way to achieve the current state of development, but much more remains to be done. In the next sections we will provide a detailed view on the state-of-the art understanding of these amazingly complex biosynthetic systems.

2. ANIMAL FATTY ACID SYNTHASE: ARCHITECTURE AND FUNCTION

2.1. Architecture of Animal FAS

2.1.1. Expression, Folding, and Assembly Elasticity. Animal FAS is a dimer formed by two intertwined chains, each with 2500 residues, and exhibits high flexibility due to its alternating links and enzymatic domains. Despite its complexity and multi-domain organization, studies have shown that the type I FAS dimer can be reversibly assembled and dissociated. In 2003, Smith et al. demonstrated that the FAS dimer, expressed in an animal host, can achieve complete dimerization after cold-induced dissociation through a simple restoration of the phosphate buffer to 0.25 M and the temperature to 20 °C.[31] A more recent study has shown that animal FAS can be truncated and reassembled to generate polyketide synthase-like modules, with evidence of domains remaining functional despite deletion of others, as well as remaining functional when isolated.[104,105]

The polyketide synthase (PKS) systems, which will be described in more detail in section 5, are responsible for the biosynthesis of various carbon skeletons—polyketides—from simple organic precursors, and also exist as both individual multi-enzymes and systems of individual enzymes.[106,107] In contrast to the barrel-shaped fungal FAS, both animal FAS and PKS megasynthases display an open architecture, with a similar domain order in the megasynthase and a similar ratio (about 10%) of insertion elements responsible for creating a structural scaffold in which domains become part of the framework. Before the determination of the first X-ray structure of animal FAS by Maier et al.,[108] several structural studies indicated that the animal FAS architecture was capable of accommodating deletion of domains without affecting the integrity of the remaining ones, as well as insertions of domains, for eventually exhibiting multi-enzymatic (type I) PKS function.[108] An example is Keatinge-Clay and Stroud's work in which they explore how domains can be swapped between type I FAS and PKS to produce a hybrid synthase.[109] A comprehensive phylogenetic analysis between PKS and FAS systems showed evidence of common ancestry and evolutionary-relatedness between the animal FAS and the type I PKS megasynthase family.[110] A subsequent study has shown that animal FAS can be truncated and reassembled to generate PKS-like modules, with evidence of domains remaining functional despite the deletion of others, as well as remaining functional when isolated.[111] Accordingly, MAT, ACP, and TE standalone enzymatic domains conserve their fold and function.[104,112] To finally cement the elasticity of type I FAS expression, folding, and assembly, both type I FAS and PKS systems have been successfully expressed in a myriad of heterologous hosts to produce mutant FASs.[113−116]

2.1.2. Overall Architecture and Catalysis. All *animal* FASs are type I systems, so we will omit this detail from now on, for simplicity. The animal FAS (EC 2.3.1.85) is a 270 kDa homodimer (Figure 1), with two identical polypeptide chains, coiled and oriented head-to-head.[117] Each monomer comprises the seven distinct catalytically active domains needed for fatty

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

acid biosynthesis: MAT, KS, DH, ER, KR, TE, and ACP. The ACP domain comprises a serine residue modified by the PNS prosthetic group,[17] and it is responsible for the substrate shuttling across all the other FAS domains.

De novo fatty acid biosynthesis in animals is a process carried out iteratively. It starts with its primer substrate, acetyl-CoA, which is repeatedly extended with malonyl building blocks, two carbons at a time, with the release of CO_2 upon condensation; the substrate is transported from one domain to another bound to the PNS arm of the ACP shuttling domain. The product is released once the chain-terminating TE domain cleaves the thioester bond between the product and ACP's prosthetic arm, which generally happens when it is 16 carbons long.

In 2008, the three-dimensional crystal structure of *Sus scrofa* (porcine) FAS was solved at 3.2 Å resolution (PDB code: 2VZ8/2VZ9).[108] Despite its groundbreaking reveal of the animal FAS structure, the latter only encapsulated five of the seven catalytic domains, lacking both ACP and TE, as well as their connecting linkers. The reason behind this is their inherent mobility which impeded finding stalled positions, and therefore, ACP and TE domains could not be traced in electron density. Ten years later, a homology model of the human type I FAS unveiled its entire structure, packing 14 catalytic domains in two monomers.[118]

Presently, there are several crystal structures of animal FAS domains available at the Protein Data Bank (PDB). These structures can be found in Table SI-1 in the Supporting Information.

2.1.3. Head-to-Tail vs Head-to-Head Orientation. Even though each of the monomers of the animal FAS homodimer possess all catalytic domains necessary for carrying out a full catalytic process (a "multi-enzyme cycle", with the production of palmitate), the monomeric form of animal FAS is unable to perform the cycle and thus to synthesize fatty acids; however, when the monomers are re-associated, the animal FAS full activity is restored. This finding was established in 1970 and, at the time, it appeared that the loss of activity through dissociation occurred at the condensation reaction.[119] This reaction requires close interaction between the PNS moiety of the ACP and the active-site cysteine of the KS domain: first, through the transfer of the saturated acyl moiety from the ACP prosthetic arm to the KS cysteine, and then in the condensation of the *S*-acyl cysteine and *S*-malonyl PNS moieties. Early biochemical finds by Stoops et al. and Salih Wakil et al. established that the cysteine residue (Cys161), present at the KS active site, could be cross-linked (using 1,3-dibromopropanone) to the thiol end of ACP PNS of the opposite subunit. These cross-linking assays suggested that catalytic domains from different monomers interact to carry out the catalysis, which was further supported by the requirement of dimerization for animal FAS to be active. Thus, the head-to-tail model emerged: the fully extended polypeptide chains of animal FAS lay side by side, in an antiparallel orientation,[120−122] creating two distinct sites for the condensation reaction to occur. The KS (residues 1−406) and ACP (residues 2125−2194) domains are far apart in the linear sequence, further corroborating the head-to-tail model (Figure 1).

However, interest had been drawn to this question. Using a mutant complementation approach,[63] in which different domains exhibited knockout mutations, it was revealed that the ACP domain interacted with KS and AT domains of both monomers (and not just the opposing one),[123] and that the DH domain cooperated with the ACP domain of the same monomer. Based on this evidence, the two animal FAS subunits could not be fully extended in a head-to-tail orientation; instead, they are likely coiled, allowing interactions between far apart domains of the same subunit. In 2004, Witkowski et al. gathered a substantial body of evidence supporting an alternative model in which the KS domains are in close proximity and play a significant role in the stabilization of the polypeptide dimerization of the complex—when devoid of KS domains, the animal FAS does not form a dimer.[117] Therefore, with a clear proven KS−KS interaction instead of the previously proposed KS−ACP interaction, it was clear that the subunits are arranged in a head-to-head fashion, as opposed to the former model.

2.1.4. Structural Scaffolding and Flexibility. Despite the unifying classification of type I FAS, which include CMN bacteria, fungi, and higher eukaryotes, its members exhibit remarkably different architectures. The animal FAS resembles a humanoid shape, with its catalytic domains exposed to the solvent, and has about 10% of its residues acting as scaffolding elements in the form of flexible linkers connecting the several domains. Additionally, no scaffold elements are found in the catalytic cores.[108] Conversely, CMN bacterial and fungal type I FAS have evolved toward a more rigid structure—with 35% and 50% of scaffolding residues, respectively[115]—where functional domains are positioned within a barrel-shaped system. This structural similarity, as well as extensive scaffolding, suggests that an evolutionary correlation between CMN bacterial and fungal FAS might exist. In a study reporting the three-dimensional cryo-electron microscopy structure of the recombinant type I FAS protein of *Mycobacterium tuberculosis* (which belongs to the CMN group bacteria), the narrow phylogenetic distribution of type I FAS systems in bacteria (only present in the genera *Corynebacterium*, *Mycobacterium*, and *Nocardia*), and the structural similarity between *M. tuberculosis* and fungal type I FAS, insinuates that an early horizontal gene-transfer event might have occurred from fungi to CMN bacteria.[115,124] Another hypothesis is that the CMN type I FAS represents an ancient type I FAS which further evolved in fungi and remained preserved in CMN bacteria[115]—interestingly, mycolic acid-producing bacteria.[125]

Comparative analysis between animal, fungal, and bacterial FAS showed that type I FAS systems are organized differently, with the main difference residing on the mechanisms of substrate shuttling during the process of fatty acid synthesis. Starting with the fungal protein (Figure SI-1 in the Supporting Information), and its impressive 50% structural scaffolding and rigid embedding of the enzymatic domains,[57,58,108,126] the ACP domain is kept inside the barrel-shaped structure of the multi-enzyme, as well as an intrinsic phosphopantetheine transferase domain (responsible for the posttranslational modification of the ACP domain).[127] In 2020, Singh et al. reported a γ subunit in yeast type I FAS inner cavity. This subunit is conserved in fungi, and it is responsible for stabilizing a rotated FAS conformation and rearranging ACP domains from equatorial to axial positions.[128] In the case of CMN bacteria, particularly *M. tuberculosis*, the overall FAS architecture is similar to that of the fungal FAS (barrel-shaped), with the difference of structural scaffolding limited to 35%. The decrease in rigidity confers remarkable structural variability. The mobile ACP is, once again, confined inside a barrel-shaped FAS scaffold. However, the former is moderately flexible, exhibiting a different position of the MAT domain,[115] which shifts to interact with the KR domain of both its subunit and the neighboring one, creating a window to the inner reaction chamber.[115] These shifts also likely affect the surface properties of the inner reaction chamber and consequently influence the residence time of ACP in the

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

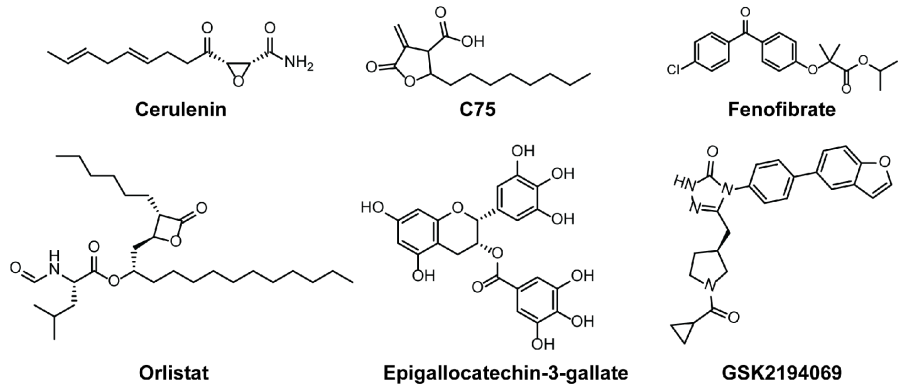


Figure 2. Structures of some of the most studied FAS inhibitors. Cerulenin, a natural product of *Cephalosporium caerulens*, and C75, a synthetic analogue of cerulenin, are two of the earliest known inhibitors of FAS targeting the KS domain and directed at cancer treatment.[166] Fenofibrate is a fibric acid derivative able to suppress FAS activity, that binds to the Ser-His-Asp catalytic triad of the TE domain. Orlistat is a synthetic drug used for the treatment of obesity, derivative of the naturally occurring lipstatin.[167] Epigallocatechin-3-gallate is a natural polyphenol present in green tea and a natural inhibitor of FAS with potential application in cancer therapy.[168] GSK2194069 is one of the most recently synthetic inhibitors of FAS, it targets the KR domain.[169]

immediacy.[129] Approximately at the same time, Ciccarelli et al. published a cryo-EM structure of the recombinant type I FAS protein of *M. tuberculosis*, Boehringer et al. published the cryo-EM structure of *Mycobacterium smegmatis* type I FAS, which does not exhibit the window to the inner reaction chamber. This may indicate that the opening may occur in CMN-bacterial FAS due to the reduced domain−domain interfaces, but the barrel is not obligatorily open.[130] The scaffolding established in animal FAS is mainly found in connecting linkers through the homomeric protein, the most important ones found at the intersection of condensing and modifying sections, as well as at the terminating section. The longest linker, approximately 20 residues, is connecting the ACP and TE mobile domains, and it has not been experimentally resolved.[108] To our knowledge this connecting linker has only been studied for the human FAS by homology modeling.[118] Through visual inspection of a classical molecular dynamics simulation of the entire human FAS protein, it was possible to observe the tremendous mobility offered by the linker.[118] Interestingly, the ACP-TE linker is remarkably plastic, withstanding modifications such as (i) an entire substitution by the longer linkers from other mega-synthases (with no effect in the activity of the enzyme) or (ii) almost complete deletion while still maintaining 30% of activity.[131]

The humanoid shape of animal type I FAS causes its catalytic domains to be exposed to the environment, with its TE domain not confined to the reaction chamber, and the ACP and the KR domains connected through a linker. Past research has shown that ACP can also access MAT and KS of the opposing subunit, which suggests an alternative mechanism for structural variability.[123] In 2009, a study identified 16 distinct FAS conformations using EM analysis. The most outstanding movement occurs at the narrow connection, i.e., the linker, between the condensing and modifying sections of the megasynthase. This "hinge" allows two distinct motions: a swiveling motion and a pendulum-like motion.[132,133] These movements, together with the length and flexibility of the tether connecting ACP to the main body of type I FAS, allow for the substrate shuttling across the type I FAS systems.

2.2. Physiological Role of Animal FAS and Its Role in Therapeutics

In a living organism, the majority of fatty acids is obtained from diet—whenever a high fat/low carbohydrate diet is ingested, FAS activity is low, and the excess fat is stored in the lipogenic tissues (liver, adipose tissue, or lactating mammary glands) in the form of triglycerides. The same occurs in the case of a low fat/high carbohydrate diet but, this time, the stored triglycerides are formed by *de novo* synthesized fatty acids due to the activity of FAS.[134] Thus, the contribution of FAS for the fatty acid metabolism is mainly dictated by the organism's diet.[135]

2.2.1. Mechanisms of Animal FAS Expression and Implications in Disease. The increased activity of FAS in a high carbohydrate diet is mediated by the production of insulin or by transcription factors that are activated by the binding of carbohydrates. In the case of insulin activation, transcription factors such as the sterol regulatory element binding protein 1c are activated by the binding of insulin to its receptor and promote expression of FAS.[136] Another type of transcription factor is activated by the high glucose concentrations in the liver, as is the case of the carbohydrate response element binding protein, also promoting FAS expression.[137] Both effects act synergistically to induce FAS expression to convert excess carbohydrates into fat. On the other hand, dietary fat has an inhibitory effect on the FAS expression, which is mainly mediated through polyunsaturated fatty acids that inhibit both of the aforementioned transcription factors.[138,139]

In a fasted state, glucagon and cAMP rule over FAS expression and are responsible for the inhibition of fatty acid synthesis, thus enabling fatty acids to undergo β-oxidation and to be used as an energy source.[140]

The above-mentioned regulatory mechanisms of the FAS act at the transcriptional level, activating or repressing the expression of the multi-enzyme in response to dietary or hormone stimuli. Nonetheless, there is evidence that the FAS activity can be rapidly modulated also at the post-translational

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

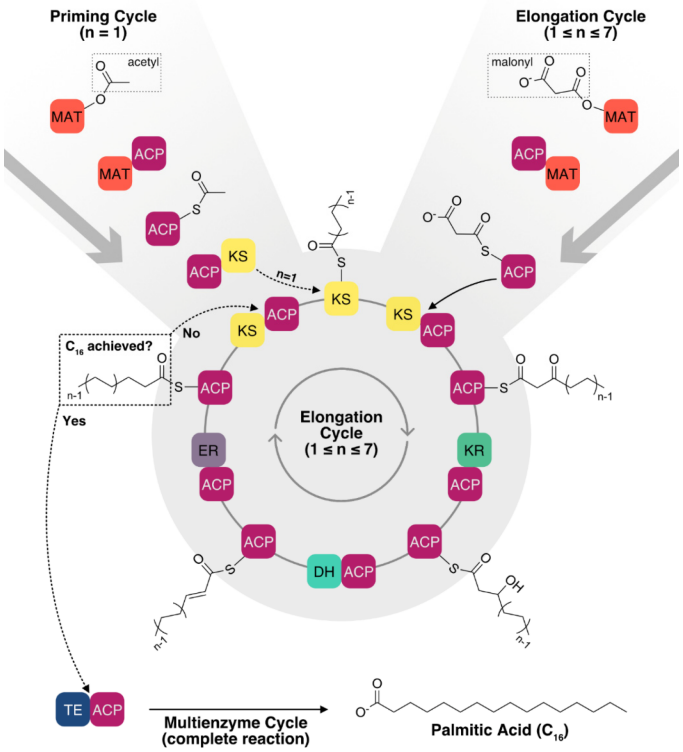


Figure 3. Proposed biosynthetic pathway catalyzed by FAS.[122] The pathway starts with the loading of acetyl-CoA (priming cycle, upper left) and continues through seven elongation cycles (upper right and center). C_{16} palmitic acid is formed as the final product of the cyclic set of reactions (the multi-enzyme cycle). For clarity purposes, the release of CO_2 and H_2O (during the KS- and DH-catalyzed reactions, respectively) and the consumption of NADPH (during the reduction steps performed by the KR and ER domains) were omitted from the figure. The iteration number (n) within the multi-enzyme cycle is shown (n = 1 for the initial cycle and n = 7 for the final cycle).

level by insulin-mediated mechanisms,[141] phosphorylation,[142] or acetylation[143] of FAS residues.

The role of the FAS in the fatty acid metabolism might surpass its ability to synthesize fatty acids, since it is also associated with the synthesis of a signaling molecule, 16:0/18:1-glycerophosphocholine,[144] which activates the peroxisome proliferator-activated receptor α (PPARα) transcription factor that triggers the transcription of genes responsible for the catabolism of fatty acids through β-oxidation.[145−148]

Since FAS plays such a central role in the metabolism of fatty acids, it has been targeted for the treaeases such as obesity. Inhibitors toward the hepatic FAS, such as the platensimycin, have demonstrated to inhibit its activity, thus reducing triglyceride accumulation in the liver, a health condition that is typically associated with obesity.[149]

FAS expression profile varies with aging. As cell proliferation rates diminish in adults, levels of the animal FAS encoding gene (*FASN*) expression are residual, and fatty acids are mainly obtained from diet.[150] However, malignant cells exhibit a lipogenic phenotype, a key hallmark in several cancer cells, derived from the overexpression of the *FASN* gene.[151−155] The up-regulation of the gene is achieved by gene amplification and through several transcriptional and translational mechanisms.[134,156−159] Its enzymatic activity can also be regulated by phosphorylation.[134] A correlation between increased lipogenesis and lower survival rates from many cancer types has been established.[160−164] The difference of FAS expression between healthy adult and cancer cells, in addition to the demonstrations that siRNA knockdowns or pharmacological inhibitors induce apoptosis of tumor cells,[81−84] has shown the potential of human FAS as a pharmaceutical target in cancer treatment.

2.2.2. Drug Design Targeting Animal FAS. As a multidomain enzyme, FAS has multiple drug targets, at least one for each of its domains, which can be explored to inhibit the whole enzyme aiming at the treatment of tumors.[165] The availability of X-ray structures of all domains of FAS enabled a structure-guided rational design of inhibitors.

Many compounds that inhibit FAS are currently known, most notably Cerulenin, C75, Orlistat, and C93, among others, like natural polyphenols such as epigallocatechin-3-gallate[168] (Figure 2). Cerulenin and C75 (which was designed to solve Cerulenin's chemical instability) are two of the earliest inhibitors of FAS that target its KS domain.[166] A crystal structure of fungal FAS from *S. cerevisiae* in complex with

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

Cerulenin (PDB code: 1B3N) shows that Cerulenin inhibits the catalytic reaction of the KS domain by forming a covalent bond between its C2-epoxide carbon and the sulfur atom of the catalytic cysteine which prevents the transfer of the acyl moiety.[170] C75 was also shown to interact with the ER and TE domains.[171] Both have exhibited promising results for cancer treatment, particularly breast and ovarian.[85,172,173] However, those come with pronounced side effects related to reduced food intake and noticeable weight loss, which is a setback for their usage on cancer therapy.[98] Therefore, the research for an inhibitor of FAS aiming for the treatment of cancer led to the coincidental discovery that FAS can also be a therapeutic target for the treatment of obesity.[99]

Orlistat, a reduced form of the natural product lipstatin, is a U.S. Food and Drug Administration (FDA)-approved drug for the inhibition of pancreatic and gastric lipase for the treatment of obesity.[174] It was first discovered on an activity-based drug screening study that revealed Orlistat's beneficial side effects as a cancer-treating drug.[83] An X-ray structure showed that Orlistat irreversibly inhibits FAS through covalent binding of its highly reactive β-lactone group to the active-site serine of the TE domain.[175] It halted tumor cell proliferation and induced tumor cell apoptosis in mice, indicating that compounds with reactive β-lactones, like Orlistat, have potential animal FAS inhibiting action.[83,167] Despite the promise of effective action in inhibiting FAS activity, Orlistat presents several limitations: chemical instability, low bioavailability, and damaging side effects such as body weight loss. This halted Orlistat's clinical development.[165,176]

Fenofibrate is another experimentally proven inhibitor of the FAS TE domain. Fibric acids are drugs effective in lowering lipid levels.[177] Fenofibrate is a fibric acid derivate, with known FAS activity suppressive qualities, and responsible for increased adipose triglyceride lipase expression.[177] Docking simulations performed with fenofibrate in the X-ray structure of the TE domain of FAS (PDB code: 2PX6)[175] have shown fenofibrate's excellent binding affinity to the active-site Ser-His-Asp catalytic triad of the TE domain of FAS, similarly to Orlistat. Furthermore, the same study demonstrated fenofibrate's cytotoxic effects in human hepatoma Hep3B cells, in a time-dependent manner, and its inhibiting power in the expression of key enzymes in fatty acid metabolism.[178]

The compound GSK2194069 was discovered in a high-throughput screening campaign using human FAS. It is a very potent inhibitor of the enzyme, specifically targeting the KR domain. Accordingly, this compound can prevent cell growth and proliferation in tumors.[169]

Another strategy that has been followed to inhibit FAS is to find inhibitors competing with the NADPH cofactor. During a screening targeting specifically small molecules that could inhibit NADPH action, by calibrating the molecule library toward their similarity to purine, a high number of molecules that inhibit FAS were identified; however, it was not possible to verify their way of action or if they indeed competed for binding with NADPH.[179]

In short, the KS, KR, and TE domains of FAS have been the main targets of small molecule inhibitors. Animal studies with recently developed molecules have been showing promising results to be clinically trialed in humans, increasing the potential of FAS to become a viable therapeutic target for the treatment of cancer, obesity, and related diseases (reviewed in refs 165, 180, and 181).

3. THE CATALYTIC MECHANISM OF ANIMAL FAS

The fatty acid biosynthesis performed by FAS, and in particular animal FAS, occurs through a cyclic pathway that comprises a recurrent set of chemical reactions, in which an initial acetyl moiety is successively extended with C_2 units derived from the elongation substrate malonyl-CoA (Figure 3).[25,122,182−184] The reaction is initiated by the MAT domain, which transfers the starter acetyl group from acetyl-CoA to the PNS cofactor of the ACP domain (Figure 3, priming cycle). This acetyl moiety is then passed to the active site of the KS domain, and the free ACP is again loaded by MAT, this time with a malonyl moiety (Figure 3, elongation cycle). Subsequently, the KS catalyzes the formation of a new C−C bond between the KS-bounded acetyl and the ACP-bound malonyl moieties through a decarboxylative Claisen condensation reaction, yielding a β-ketoacyl-ACP intermediate. The β-carbon of the condensation product is then processed by a NADPH-dependent reduction by the KR domain, followed by a dehydration reaction catalyzed by the DH domain and an additional NADPH-dependent reduction catalyzed by the ER domain, leading to a four-carbon saturated acyl-ACP substrate. This newly formed saturated ACP-bounded substrate is translocated back to the KS domain and, in a second elongation cycle, is condensed with a new malonyl moiety and proceeds through the same three-step β-carbon processing reactions, resulting in a saturated acyl-ACP substrate that is two carbon units longer than in the preceding iteration. The biosynthetic cycle continues until a saturated C_{16} palmitoyl thioester is achieved (the final product of seven elongation cycles), which is cleaved from ACP by the thioesterase (TE) domain to give a full-grown saturated free fatty acid (Figure 3, multi-enzyme cycle). TE might also cleave the ACP-bound intermediate before completing the seven elongation cycles, or after further cycles, but the C_{16} palmitic acid product is vastly dominant.[25] The set of reactions performed by each FAS domain that result in the formation of a palmitic acid, as well as the fundamental elements of the ACP path along the biosynthetic pathway, can be better understood by watching Video SI-1 in the Supporting Information.

3.1. Reaction Cycle of the MAT Domain

3.1.1. Substrate Binding and Promiscuity. Acyl moieties are used as the C_2 building blocks in the fatty acid synthesis. These moieties are first chemically bound to the PNS cofactor of the ACP domain, one by one, and each pair of carbon atoms is subsequently presented to the FAS catalytic domains that reduce them to aliphatic carbons until a full-grown fatty acid is produced. The acyl substrates, originally a part of acetyl-CoA or malonyl-CoA, are received and translocated to the ACP domains by enzymatic acyltransferase (AT) domains, present in all FAS systems.[18,185−189]

The fungal/bacterial FAS exhibit two different AT domains/enzymes. One of them transfers the primer acyl substrate, and the other transfers the extender acyl substrates to the ACP.[190−197] The AT domain of animal FAS, MAT, is more versatile, being capable of loading both the acetyl starter and malonyl extender molecules.[187,188,198,199] Past studies have reported that the acetyl-CoA and malonyl-CoA substrates randomly bind to a common active site of the FAS (i.e., the MAT domain) which, consequently, makes them competitive inhibitors of each other's binding.[122,187,200−203] Thus, for an efficient fatty acid synthesis, the inappropriately bound acetyl or malonyl moiety must be rapidly unloaded (i.e., the acyl-FAS bond must be hydrolyzed) and transferred back to the CoA until

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

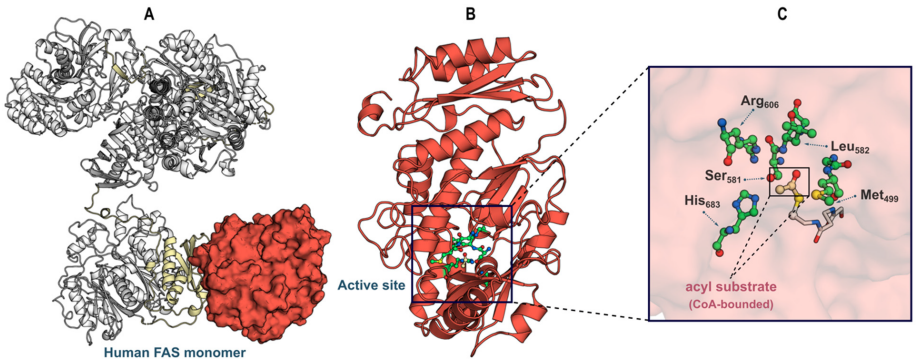


Figure 4. The malonyl-acetyl transferase (MAT) domain of animal (human) FAS. (A) Three-dimensional surface view of the MAT domain within the human FAS monomer. (B) Cartoon representation that evidences MAT's secondary structure elements and active-site region (black square). This MAT-substrate complex corresponds to the model used in a recent computational study[223] and is based on the crystallographic structure of the human MAT domain (PDB ID: 2JFK).[221] (C) Close-up of the active-site cavity. The catalytic residues (human FAS numbering) are depicted in ball-and-stick representation, and the acyl substrate, that should be bound to a CoA molecule (not shown for simplicity), is highlighted by the black rectangle.

a competent enzyme complex is formed.[199,204] The dual specificity of MAT makes it responsible for initiating the fatty acid synthesis reaction and for continuously providing the C_2 extender substrate (i.e., malonyl moiety) at each elongation cycle. It is noteworthy that, in rare circumstances or in certain specialized tissues, the MAT domain is capable of accepting unusual CoA-esters substrates such as propionyl-, methylmalonyl-, and phenylacetyl-CoA, and of incorporating them in the fatty acid synthesis. This ultimately leads to the production of odd-numbered or even-numbered branched fatty acids,[104,205−209] but at a much lower extent than the synthesis of the predominant C_{16} fatty acid palmitate.

3.1.2. Active-Site Structure and Catalytic Mechanism. Over the past decades, numerous studies have explored the catalytic machinery of the MAT domain (Figure 4A) and its fungal and bacterial homologues. The experiments were made in a large number of different animals, fungi, and bacteria. It is implicitly assumed that the results obtained in one organism can be generalized to other organisms within the same kingdom, as far as the same type of FAS is present. For example, experiments in animal FAS have been carried out with protein from pig, pigeon, goat, human, or rat, among others.

In 1970, Phillips and co-workers reported that a hydroxyl site of the animal (pigeon) FAS transfers the CoA acetyl and malonyl moieties to a sulfhydryl acceptor (i.e., the PNS group of the ACP).[202] Later that year, Joshi et al. tentatively identified the hydroxyl site as the hydroxyl group of a serine residue,[200] similarly to what was reported for the bacterial (*E. coli*) malonyltransferase as well as for the fungal (yeast) malonyltransferase domain.[193,210] Moreover, the observation that acetyl and malonyl groups bound to the hydroxyl site of the animal (pigeon) FAS and formed stable intermediates indicates that the acyl moieties are, in a first stage, transferred to the catalytic serine of the MAT domain of FAS and, in a second stage, loaded into the ACP domain.[200,202] This serine-centered, two-stage transfer reaction is consistent with a ping-pong bi-bi mechanism[211] that is shared among the AT enzymes/domains of all FAS types.[47,186,191,193,212−215]

Peptide mapping studies reported that the catalytic serine of the MAT occupies the position 581 in the animal (goat) FAS (unless otherwise stated the residue numbering always corresponds to the homologous human enzyme, UniProtKB: P49327), located within a highly conserved G-X-S-Y-G sequence that lies at a sharp turn between an α-helix and a β-strand (Figure 4B).[201] This structural motif, commonly referred to as the "nucleophilic elbow",[216,217] is shared between all members of the α/β-hydrolase superfamily, which are known to resort to an acid−base−nucleophile triad to catalyze their reactions.[218] As a reminiscent of such triads, the ATs involved in the fatty acid synthesis retain a Ser-His catalytic dyad, where the serine acts as the nucleophile and the histidine residue as the base that deprotonates the serine, enhancing its nucleophilicity toward the acyl substrate.[37,219] Crystallographic studies performed with the bacterial (*E. coli*) malonyl-CoA-ACP transacylase enzyme (named FabD), loaded with the extender malonyl-CoA, reported the existence of a hydrogen bond between the active-site residues Ser92 and His201,[37,220] which were later found to align with the animal (human and rat) FAS Ser581 and His683.[104,219,221] By performing a set of mutagenesis experiments, Rangan and co-workers were the first to demonstrate the implication of the Ser581-His683 catalytic dyad (Figure 4C), which seems to be present in all ATs, in the catalytic mechanism of the animal (rat) MAT domain.[219] The replacement of the active-site Ser581 by an alanine residue completely inactivated MAT toward its acetyl- and malonyl-CoA substrates, thus validating its crucial role as the nucleophilic species. Equivalent mutations performed on the bacterial (*Streptomyces coelicolor*) ATs produced the same outcome, as their ability to transfer acyl moieties was completely abolished,[215,222] and confirmed the existence of a common chemical mechanism for all of them. Regarding His683, its replacement by an alanine reduced both the transacylase activity and the rate of acylation of the nucleophilic Ser581 by more than 3 orders of magnitude. This demonstrated that His683 is essential for the overall reaction, where it should act as the main residue responsible for the activation of Ser581 by deprotonating the serine's hydroxyl group.[219]

The ability to recognize and transfer both acetyl and malonyl moieties with similar efficiencies is a singular characteristic of the animal MAT domain, as aforementioned. A set of multiple

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

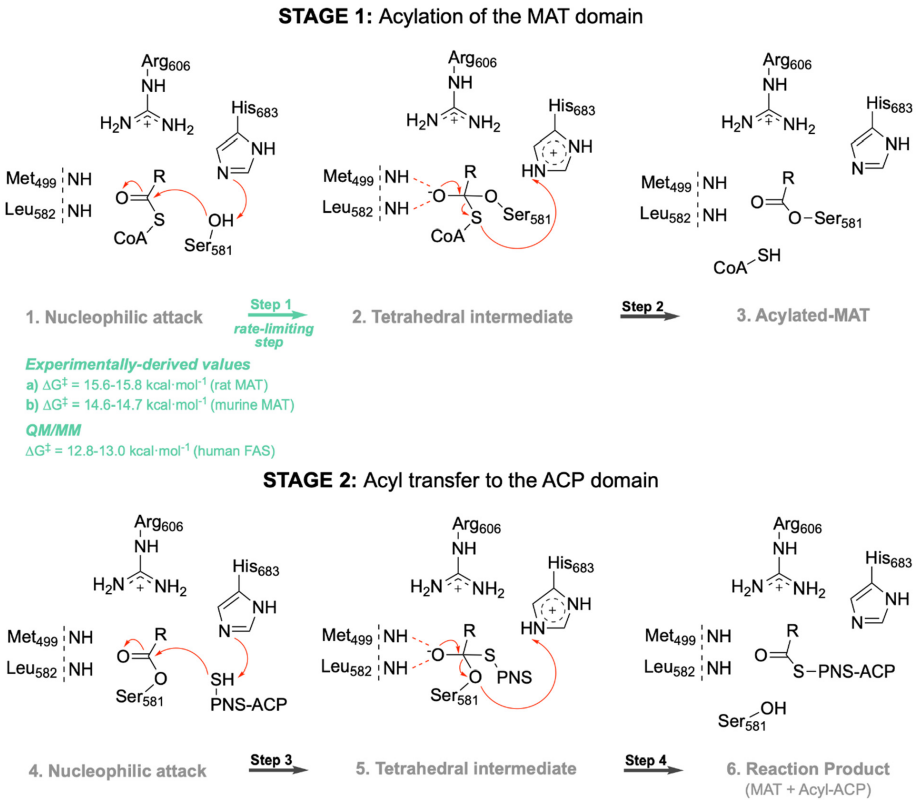


Figure 5. Catalytic mechanism of the animal FAS MAT domain ("R" stands for acetyl or malonyl moieties). This ping-pong bi-bi mechanism proceeds through two major stages, where in the first an acyl moiety is transferred from CoA to a serine in the MAT domain and, in the second, the same acyl molecule is loaded onto the thiol of the PNS cofactor of an ACP domain. The Gibbs activation energies of 15.6−15.8 kcal·mol^{-1} and 14.6−14.7 kcal·mol^{-1} were derived from the k_{cat} values of 1.2−1.9 s^{-1} (0 °C)[224] and 99.2−119 s^{-1} (25 °C),[104] respectively, using transition state theory (TST).[227] The Gibbs activation energies of 12.8−13.0 kcal·mol^{-1} were determined using QM/MM methods.[223,226]

sequence alignments revealed that the basic residue Arg606 (Figure 4C) is universally conserved among the (M)ATs involved in fatty acid synthesis and also in specific ATs of PKS modules.[219] Since all of these (M)ATs recognize malonyl moieties, it was hypothesized that this basic residue plays an important role in the malonyl binding, particularly by interacting with the carboxyl group of the substrate. Indeed, the mutation of Arg606 by an alanine residue severely impairs the MAT's ability to transfer malonyl moieties without compromising the efficiency of the acetyl transfer.[104,224] Furthermore, the analysis of X-ray structures of malonyl-AT complexes in the bacterial (*E. coli*) FabD enzyme,[220] the homologous PKS DynE8s AT domain,[225] and the animal (murine) FAS MAT-KS didomain[104] evidenced the existence of bidentate ionic bonds between the guanidinium group of the conserved arginine and the malonyl carboxyl group, which may assume a side-on or an end-on character. Molecular docking studies performed with the human MAT have also detected these Arg606-malonyl salt bridges,[221] and their contribution for the malonyl transfer was also validated by Quantum Mechanics/Molecular Mechanics (QM/MM) calculations.[223]

The capacity for binding acetyl-CoA can be explained by the more hydrophobic environment of the active site of MAT when compared to that of the bacterial and PKS malonyl-specific transferases.[108,221]

Throughout the whole MAT-catalyzed reaction, the methyl group of the acetyl substrate remains lodged in a hydrophobic pocket formed by Met499-Phe553-Phe682, whereas polar residues usually form this region in most PKS and bacterial homologues.[108,223,225,226] When an alanine replaces Arg606, both the vacant volume and the hydrophobic character of the active-site increase, thus explaining why the mutant enzyme is able to handle longer-chain acyl moieties and exhibits enhanced activity toward acetyl-CoA, respectively.[221,223,224] The hydrophobic nature of the MAT active site renders the acetyl- and malonyl-enzyme intermediates less susceptible to the solvent-mediated hydrolysis, which explains why the transacylation to the ACP domain occurs about 3−4 orders of magnitude faster than hydrolysis.[104,225]

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

Besides the participation of the above-mentioned catalytic Ser-His dyad, the substrate-binding characteristics of the Arg606 residue, and the hydrophobic nature of the active site (Figure 4C), further molecular characteristics are essential to promote the MAT-catalyzed reaction. Numerous theoretical and experimental studies have unveiled their nature and importance, allowing the establishment of the catalytic mechanism of the animal MAT domain with atomistic detail, represented in Figure 5. The following subsections discuss the most relevant catalytic aspects of the two-stage, MAT-catalyzed reaction.

STAGE 1: Acylation of the Animal MAT Domain. In the first catalytic stage, the MAT domain is acylated as the result of a nucleophilic attack performed by the deprotonated Ser581 on the substrate thioester carbonyl carbon (Figure 5, STAGE 1). A recent theoretical work that employed QM/MM methods has provided insight regarding the succession of events and energetics of this particular stage.[223] While it was thought that the deprotonation of the hydroxyl group of Ser581 by His683 preceded the nucleophilic attack, the computational calculations ruled out that hypothesis, as this rendered an energetic barrier substantially higher than that experimentally derived for the transfer of either acetyl or malonyl (~14.6−15.8 kcal·mol^{-1}, using transition state theory (TST)).[104,224,227] Instead, upon the binding of the substrate, the attacking hydroxyl oxygen of Ser581 binds the acyl carbonyl carbon and, concertedly, the Ser581 is deprotonated by His683 (Figure 5, STAGE 1). At the transition state, the proton is almost completely transferred to the His683 N_ε, while the acyl-Ser581 oxo-ester bond is still under formation. The computational calculations revealed that this concerted step is rate-limiting for the whole human MAT-catalyzed reaction,[226] with a Gibbs energy barrier of 13.0 or 10.9 kcal·mol^{-1} for acetyl-CoA and malonyl-CoA, respectively. A tetrahedral intermediate is formed at the end of the concerted nucleophilic attack, as typically occurs in ATs with a serine nucleophile.[37,220,221,228] In this intermediate structure, the thioester bond between CoA and the acyl moiety is practically disrupted, and the bond between Ser581 and the acyl moiety is nearly formed, thus forcing it to adopt a tetrahedral geometry.

The geometry of the "nucleophilic elbow" contributes to the formation of an oxyanion hole which stabilizes the transition state and tetrahedral intermediates of this reaction. Based on the analysis of the X-ray structure of the animal (human) MAT domain, Bunkoczi and colleagues identified the backbone amides of Met499 and Leu582 as the entities forming the oxyanion hole,[221] a fact later validated in the crystallographic structure of the murine malonylated-MAT-KS di-domain.[104] The existence of strong charge−dipole interactions between these backbone amides and the partially ionic carbonyl group of the acyl moiety was found to reduce the activation energy of the concerted nucleophilic attack by ~4−7 kcal·mol^{-1}.[223] The contribution of oxyanion holes for the stabilization of transition-states has also been investigated in the context of other enzymatic systems, and similar conclusions were obtained.[229−231] In this particular case, by driving the electron density away from the carbonyl carbon to the carbonyl oxygen, the oxyanion hole makes the labile thioester bond more susceptible to be attacked by Ser581. In addition, the oxyanion hole is also accountable for the electrostatic stabilization of the tetrahedral intermediate,[223] which results from the transformation of the initial dipole−dipole interactions established with the substrate's carbonyl oxygen into stronger ion−dipole ones at the intermediate state.

The second reaction step consists of the breakdown of the tetrahedral intermediate and the consequent formation of an oxo-ester bond between Ser581 and the acyl carbonyl carbon, which is triggered by the deprotonation of the positive His683 by the CoA thiolate. At the end of this reaction step, a stable acyl-MAT complex is formed (Figure 5, Step 2). The free CoA leaves the active site, making room to accommodate the PNS cofactor of the ACP domain, which will receive the acyl moiety on a subsequent stage.

It is believed that the stability of this acyl-enzyme complex is partially due to a characteristic that is common to type I and type II FAS systems: the unique location of the oxyanion hole.[25,196,228] By blocking the positioning of a potentially hydrolytic water in the active-site cavity, the oxyanion hole prevents the hydrolysis of the bond formed between the catalytic serine and the acyl moiety. Together with the highly hydrophobic nature of the active site, this fact explains why the rate of hydrolysis is almost negligible compared to the rate of transacylation.[104,219]

STAGE 2: Transfer of Acyl Moieties to the ACP Domain. In the second stage of the MAT catalytic mechanism (the "pong"), the acyl moiety bound to Ser581 is transferred to the PNS cofactor of ACP, generating an acyl-ACP complex, and the catalytic Ser-His dyad is regenerated to its original state (Figure 5, STAGE 2).[25] This stage is very similar to the first one in what concerns the number of catalytic steps and chemical reactions, even though the chemistry takes place in the reverse order. The third reaction step takes place after the binding of ACP to the acylated-MAT, when PNS enters the active site and its thiol faces the conserved His683 and the acyl carbonyl carbon.[226] It consists of a concerted reaction that combines the His683-mediated deprotonation of the PNS thiol group and the nucleophilic attack of the thiolate on the Ser581-acyl bond (Figure 5, Step 3). The labile transition state leads to the formation of a second tetrahedral intermediate, which is stabilized, by ~7−11 kcal·mol^{-1}, through interactions performed with the backbone amides that form the oxyanion hole.

To conclude the transacylation reaction, the tetrahedral intermediate should collapse to yield a regenerated MAT and an acyl-ACP complex.[25,228] Indeed, computational QM/MM calculations revealed that the fourth and last reaction step (Figure 5, Step 4) occurs in a single concerted step, where the Ser581-acyl bond breaks down and PNS binds to the acyl group, and the positive His683 protonates the active-site Ser581 to restore the active site of MAT.[226] Throughout this catalytic step, the interactions between the acyl carbonyl group and the backbone amides of Met499/Leu582 become progressively weaker as the oxyanion returns to its regular carbonyl oxygen state, which eventually leads to the disassembly of the oxyanion hole.

At the end of STAGE 2, ACP leaves the MAT active site and transports the PNS-loaded acyl moiety to the KS domain, so that the fatty acid biosynthesis successfully pursues. The driving forces behind this transport, if any, and the pathway for the transport, are unknown at the moment.

3.2. Reaction Cycle of the KS Domain

3.2.1. Structure and Function of the Animal KS Domain. During the process of biosynthesis of fatty acids, an acyl precursor is successively elongated by consecutive additions of a C_2 extender. Its extension involves C−C bond-forming condensation reactions catalyzed by the FAS enzymes.[25,31,122,232,233] All type I FAS systems involved in

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

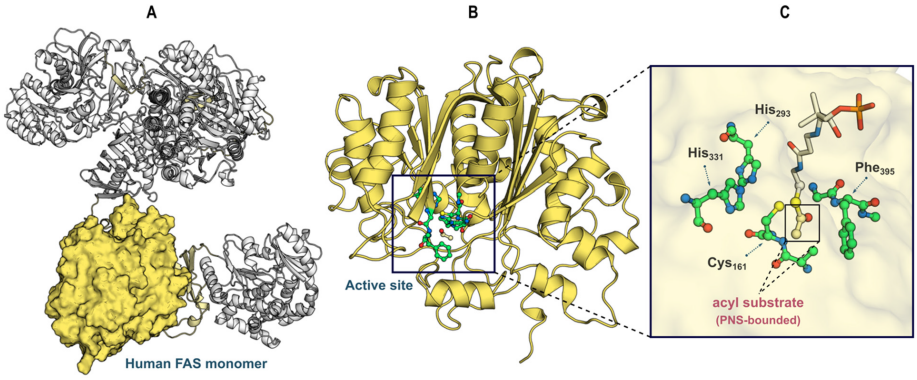


Figure 6. β-Ketoacyl synthase (KS) domain of the animal (human) FAS. (A) Three-dimensional surface view of the KS domain within the FAS monomer. (B) Cartoon representation of the KS secondary structure elements and active-site region (black square). This KS−substrate complex is based on the crystallographic structure of the human KS-MAT didomain (PDB ID: 3HHD),[245] which was modeled by the authors of this Review to add the acyl-PNS substrate. (C) Close-up of the active-site region. The catalytic residues are pictured in ball-and-stick representation, and the acyl substrate that is bound to the PNS cofactor of ACP (not shown for simplicity) is highlighted by the black square.

cytoplasmatic *de novo* synthesis rely on a single type of condensing enzyme to perform all chain-elongation steps: the KS (Figure 6A).[47,184,234,235] In the type II FAS system, multiple forms of condensing enzymes, with different substrate specificities, have been described, such as the β-ketoacyl ACP synthases I (KAS I or FabB), II (KAS II or FabF), and III (KAS III or FabH).[29,196,232,236−239] As such, the wide range of substrate specificities among these condensing enzymes gives rise to a panoply of fatty acid products formed in the respective biosynthetic pathways. Despite that, the majority of the KSs involved in the biosynthesis of fatty acids and polyketides share extensive sequence similarity and are related in both evolutionary and structural terms.[196,232,240]

It was not until the late 1990s and early 2000s that the first crystallographic structures of condensing enzymes involved in the fatty acid or polyketide synthesis were resolved.[33−35,108,241−247] In the animal FAS dimer, the two KS lie close together near the center of the multi-enzyme, Figure 1, functioning as a homodimer that plays an essential role for the stabilization of the dimeric form of FAS.[108,117,118] Structure-wise, the KS domain of animal FAS exhibits the well-known α-β-α-β-α architecture (Figure 6B), typical of the thiolase superfamily, and shares structural resemblances with its PKS, fungal, and bacterial homologues.[108] Located at the dimer interface, the substrate-binding cavity leads to the entrance of the active site that harbors the catalytic Cys161 and two conserved histidine residues, His293 and His331, forming a catalytic triad (Figure 6C), and the conserved residue Lys326.[25,245,248] Similarly to the catalytic Ser581 of the MAT domain, the Cys161 is positioned within a "nucleophilic elbow", where it experiences a positive electrostatic potential generated by the attached helix's dipole moment.[249] The occurrence of catalytic Cys-His-His triads is a common feature among the bacterial β-ketoacyl ACP synthases I and II and the KS of modular PKSs, which suggests, once again, that all these systems share identical reaction mechanisms. Furthermore, multiple experimental works demonstrated the importance of these active-site residues for the elongation reaction and for the overall fatty acid synthesis.[32,35,235,250−253]

During the first extension cycle (C_2 insertion/complete reduction) of the animal FAS, the KS domain performs a Claisen decarboxylating condensation between an acetyl moiety (the initiation primer) and a malonyl molecule (the C_2 extender substrate), each one delivered in turn to KS by ACP.[25,31,122,183,184,254,255] Given that the two FAS chains exhibit a head-to-head arrangement, the KS domains can interact functionally with the ACP domains of either chain, i.e., bind any of the two ACP and receive the PNS-bound intermediate to elongate.[31,66,108,117,256,257] However, the KS-mediated exchange of acyl intermediates between ACP domains of opposite chains during successive elongation cycles seems to be an optional event.[123]

3.2.2. Catalytic Machinery of the KS Domain. The works of Porter et al. and Wakil et al. in the early 1950s, using animal (pigeon) liver extracts, indicated that the presence of essential thiol groups was a requirement for the fatty acid synthesis,[258,259] which was later verified in animal (rat brain and pigeon liver) extracts,[260,261] as well as in bacteria (*E. coli*)[262] and fungi (fresh baker's yeast), where they specifically participate in the condensation reaction of fatty acid synthesis.[263] Finally, studies on purified bacterial (*E. coli*) β-ketoacyl ACP synthase (the bacterial analogue to the animal KS domain) revealed that this enzyme contains a functionally significant thiol group that can be protected against iodoacetamide or *N*-ethyl-maleimide inhibition by acetyl-ACP and other acyl-ACP derivatives, but not by malonyl-ACP.[264,265] This would suggest that the thiol group is involved in the acyl-ACP binding site of β-ketoacyl ACP synthase and that these protecting acyl-ACP compounds could interact with the enzyme by forming an acyl-S-enzyme state. Thus, it was proposed that the condensation reaction occurs in a stepwise manner, where the β-ketoacyl ACP synthase first accepts an acyl moiety (initially an acetyl one) from ACP to form an acyl-enzyme intermediate, and then condenses it with malonyl-ACP, leading to the formation of a β-ketoacyl-ACP complex.[264−266]

The reports of a similar thiol-protecting effect by acetyl-CoA or shorter chain saturated acyl-CoA's (but not by malonyl-CoA) in the fungi (yeast) and animal (pigeon liver) FAS have led to

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

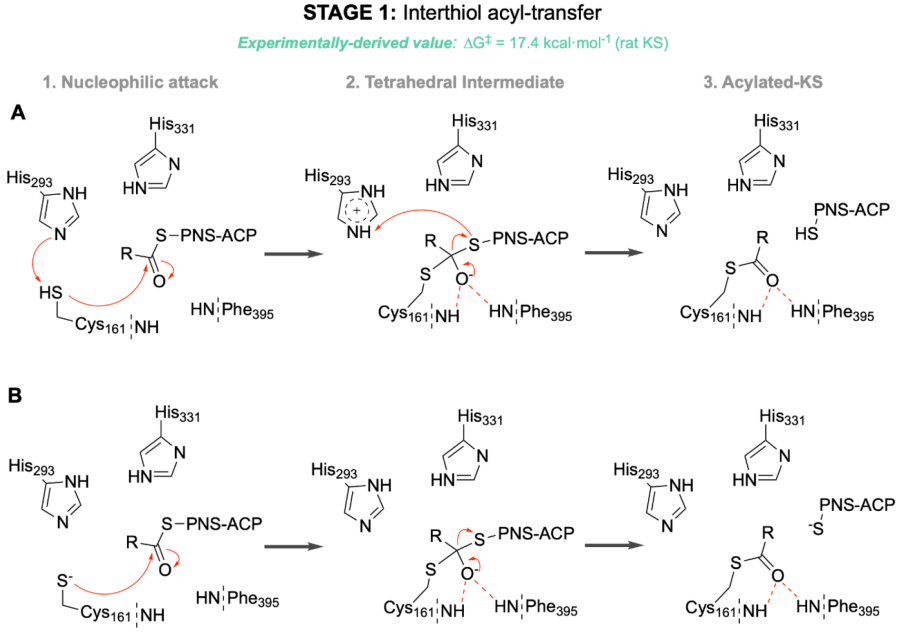


Figure 7. Proposed mechanistic pathways for the first stage (inter-thiol acyl-transfer) of the KS domain catalytic mechanism. ("R" stands for the acyl moiety that will be elongated by the KS domain. The residues are numbered according to the human KS domain). (A) According to this proposal, a stable zwitterionic Cys-His pair is formed as a result of a conformational change that occurs when the substrate enters the binding pocket. Then, the negative Cys161 performs a nucleophilic attack on the carbonyl group of the acyl moiety, and a tetrahedral intermediate is formed. Afterward, the positive His293 residue is expected to protonate the leaving PNS group. (B) This pathway assumes that, by experiencing a positive dipole-moment from the N-terminus α-helix where it is located, the Cys161 exhibits an enhanced nucleophilicity toward the carbonyl carbon of the substrate. In this proposal, no particular chemical entity is specified as the proton donor that should neutralize the PNS group after the nucleophilic attack. The Gibbs activation energy of 17.4 kcal·mol^{-1} was derived from the k_{cat} value of 1.14 s^{-1} (25 °C)[252] using TST.

the conclusion that the condensing reactions of these systems should be identical to the bacterial ones and follow a similar ping-pong mechanism.[261,263,266−269] Indeed, it was later demonstrated that the reactive thiol binding site of type I and type II KS corresponds to a conserved cysteine residue, identified as Cys161 in the animal (rat) FAS KS[64,184] (we will use the residue numbering of the human KS domain, for simplicity, even when referring to bacterial/fungal FAS systems), to where the acetyl or saturated acyl moieties should bind prior to their condensation with malonyl-ACP in each elongation cycle.[120,200,202,266,267] Only in 1997, Witkowski and colleagues demonstrated that KS possesses the intrinsic ability to perform both the inter-thiol transfer and the condensation reactions, and that substrate specificity of the condensing reaction is mainly driven by the substrate specificity of the acyl-transfer activity toward ACP and KS.[235] As KS exhibits a barely noticeable inter-thiol transfer activity toward unreduced acyl-ACP intermediates (i.e., it does not load unreduced intermediates accidentally brought by ACP), it is ensured that only the saturated acyl intermediates that have successfully completed the previous priming/elongation cycle (i.e., been sequentially modified by the KS-KR-DH-ER domains to yield a fully saturated chain) bind to the active-site cysteine and undergo the set of condensations that culminate with the formation of a fully saturated long-chain fatty acid.

KS mutants where the catalytic Cys161 is replaced by either Ser, Thr, or Ala were observed to severely impair the rate of condensation and, consequently, abolish the global activity of both type I and type II FAS systems.[32,35,235,250,251,253] As for the two His residues, the mutagenesis of these conserved residues by Ala was found to significantly reduce the ability of FAS to catalyze the condensation reaction.[32,35,252,253] In addition, a 3-fold reduction of the condensation activity of the mutant animal (rat) FAS was reported by Witkowski et al. as a result of the simultaneous replacement of both His293 and His331,[252] thus emphasizing their importance for the elongation reaction.

Despite the existence of numerous studies focused on the KS-mediated decarboxylative Claisen condensation, the aspects concerning the specific role of the active-site residues and the sequence of catalytic events remain poorly understood and a matter of controversy (the authors of this Review are actively working to shed light on this issue). It is presently accepted that the domain reaction cycle catalyzed by animal KS, and its bacterial and fungal homologues, proceeds through three sequential stages (Figure 7 and Figure 8), which will be reviewed in detail in the following subsections: (1) inter-thiol transfer of an acyl moiety from the ACP thioester to the KS catalytic Cys161; (2) binding of malonyl-ACP (previously loaded by the MAT domain) to the active-site KS and subsequent decarboxylation of the chain-extender malonyl

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

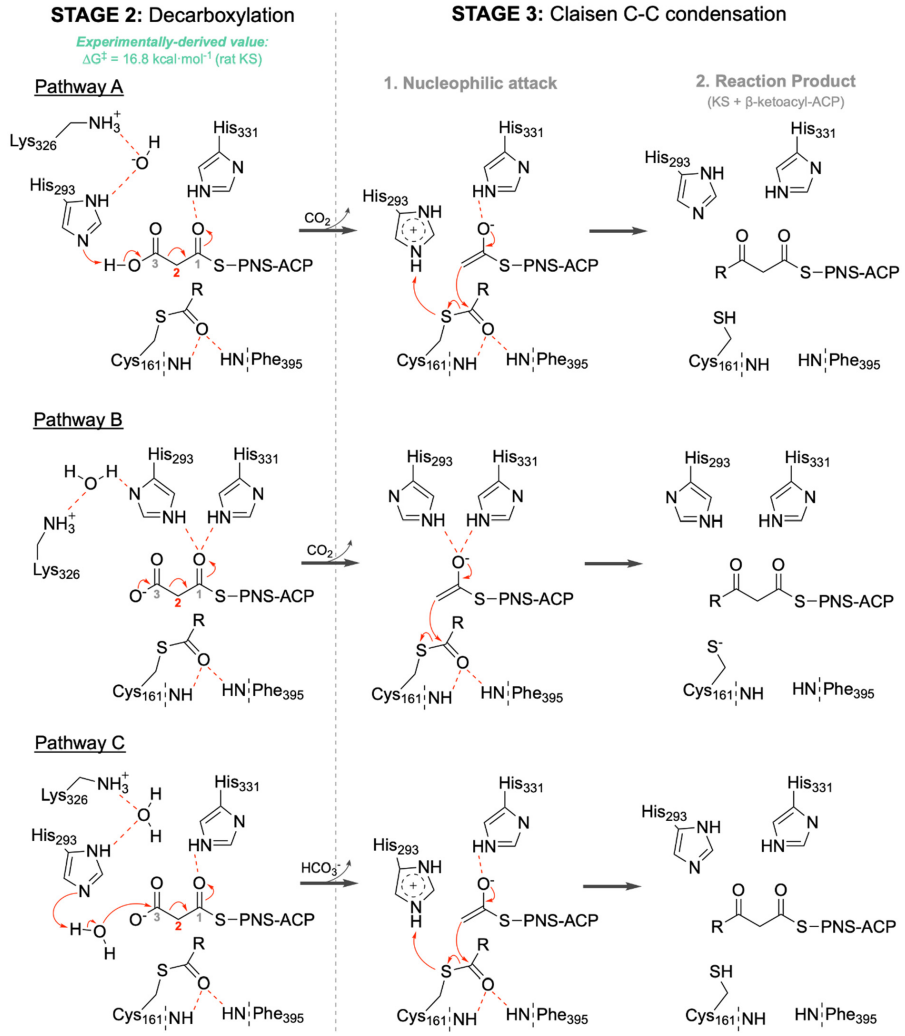


Figure 8. Proposed mechanistic pathways for the second and third stages of the KS domain catalytic mechanism ("R" stands for the acyl moiety under elongation by the KS domain). The Gibbs activation energy of 16.8 kcal·mol^{-1} was derived from the k_{cat} value of 2.82 s^{-1} (25 °C)[252] using TST.

moiety, generating a reactive intermediate with carbanion character; and (3) Claisen C−C condensation of the decarboxylated malonyl moiety with the saturated acyl moiety, yielding the β-ketoacyl-ACP product.[25,232,252,253]

STAGE 1: Inter-thiol Acyl Transfer between ACP and the KS Catalytic Cysteine. The first catalytic stage of the decarboxylating condensation reaction consists of an acyl-transfer from the PNS cofactor of one of the ACP domains to the thiol group of the KS active-site cysteine. If the fatty acid synthesis is at the priming cycle, the acyl moiety transferred to KS typically corresponds to acetyl, which is the primer substrate. On the remaining elongation cycle of the overall multi-enzyme cycle, ACP carries and transacylates KS with saturated acyl intermediates that are formed at the end of the previous priming/elongation cycle (Figure 3). Kinetic studies concerning the inter-thiol acyl-transfer activity of the animal (rat) KS domain revealed that the k_{cat} of this particular stage is 1.14 s^{-1} (25 °C), which under TST corresponds to a Gibbs activation energy of 17.4 kcal·mol^{-1}.[227,252]

After the successful binding of acyl-ACP on the active site of KS, the catalytic cysteine of KS performs a nucleophilic attack on the acyl's carbonyl carbon (Figure 7).[35,196,243,252,253] It was initially supposed that one of the two adjacent hydrogen-bonding His residues was responsible for the Cys161-thiol

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

deprotonation, acting as a general base, as for His683 at the MAT domain,[240,270] coherently with the analysis of the crystallographic structures of type II FAS condensing enzymes.[33−35,243,246] This type of His-mediated activation of a Cys nucleophile is also typical of cysteine proteases, which include the main protease of SARS-CoV-2, the etiological agent responsible for the COVID-19 pandemic.[271−274] A recent theoretical study of the *M. tuberculosis* KAS I (KasA), a bacterial homologue of the animal KS, analyzed the molecular interactions responsible for the enhanced nucleophilicity of the catalytic cysteine, and supported the deprotonation hypothesis.[275] Lee and Engels suggested that the entrance of the substrate hydrophobic tail in the binding pocket induces a conformational change in a phenylalanine residue that favors the formation of a stable zwitterionic Cys^{-}-His^{+} [$Cys161^{-}$-$His293^{+}$] state (Figure 7A). Upon substrate binding, the hydrogen bond network to the Cys161-thiolate is dismantled, and the nucleophilicity of Cys161 toward the carbonyl group of the acyl moiety increases. Surprisingly, the individual His293Ala and His331Ala mutations in the animal (rat) FAS (or the positionally equivalent histidine residues in the *E. coli* FabB and FabH homologue enzymes) have not impaired neither the reactivity of Cys161 nor the inter-thiol acyltransferase activity of the KS domain, but rather enhanced them.[32,35,252] This implies that, although these conserved basic residues may be important for the elongation reaction, their role does not appear to be related to the activation of the catalytic Cys161. Instead, it is possible that any of the two His residues may deprotonate the active-site Cys, and thus only a double mutant can impair the catalytic activity.

Alternatively, Davies and colleagues proposed that the enhanced nucleophilicity of the Cys161 could result from a pK_a decrease upon interaction with the positive dipole-moment from the long *N*-terminus α-helix where it is located[35] (Figure 7B). Later studies performed with both type I/II KS domains/enzymes provided results that are consistent with this hypothesis, currently regarded to be more plausible.[32,243,248,252,253]

Throughout the first catalytic stage, a negatively charged tetrahedral intermediate is formed between the catalytic Cys161 and the thioester carbonyl of the acyl-ACP complex (Figure 7). Following a similar trend to the catalytic mechanism of the MAT domain, it is assumed that the backbone amides of two adjacent residues form an oxyanion hole that stabilizes the increasingly negative charge of the acyl moiety's carbonyl oxygen.[34,35,232,243,244,246,252] In the bacterial β-ketoacyl ACP synthases I, II, and III enzymes, the backbone amides of the active-site Cys and either a conserved Phe or a Gly residue have been identified as responsible for the stabilization of the tetrahedral intermediate.[34,35,243] Based on the highly conserved nature of the active site of type I and type II KS, Witkowski et al. tentatively identified the backbone amides of Phe395 and of the catalytic Cys161 as the residues comprising the oxyanion hole of the animal (rat) KS domain,[252] which was recently corroborated by Rittner and co-workers.[248]

Altogether, the combination of the helix-dipole-induced nucleophilicity of the Cys161 and the oxyanion hole-stabilizing effect should provide adequate energetic conditions for the inter-thiol transacylation. At the end of the first catalytic stage, the KS domain rests in the acyl-enzyme state and the KS:ACP complex dissociates. The details concerning the protonation of the thiolate of the leaving PNS cofactor are still not fully understood and depend on the activation process of the catalytic cysteine. In the zwitterionic Cys^{-}-His^{+} model (Figure 7A), the positive histidine residue is expected to donate a proton to the PNS thiolate, neutralizing the leaving group.[275] On the other hand, the helix dipole-moment effect proposal (Figure 7B) does not specify a particular chemical entity that may protonate the negative thiol of the leaving PNS group, thus requiring further studies to clarify this protonation event. In this particular case, protonation by the solvent after KS:ACP dissociation is a likely event.

STAGE 2: Decarboxylation of the Chain-Extender Malonyl-ACP. Earlier studies demonstrated that the decarboxylating event can be uncoupled from the C−C bond formation and that it precedes the latter.[251,252,268,276] In addition, Witkowski and colleagues revealed that the decarboxylation reaction is not performed unless a saturated acyl moiety is bound to the cysteine nucleophile of the KS domain, Cys161,[251] which means that the decarboxylation follows the inter-thiol acyl transfer (STAGE 1). This may represent a control mechanism that prevents the accidental loading of KS with a partially reduced acyl-ACP intermediate that has not completed a full elongation cycle. Hence, at a second stage of the KS-catalyzed reaction, the chain-extender substrate malonyl is brought by ACP to KS. ACP binds to the acylated KS domain, and the malonyl extender is subsequently decarboxylated, leading to the formation of an enolate anion (Figure 8, STAGE 2). Witkowski et al. analyzed the kinetic parameters associated with the animal (rat) KS decarboxylation reaction and observed that the k_{cat} associated with the malonyl decarboxylase activity is 2.82 s^{-1} (25 °C), which is predicted from TST to correspond to a Gibbs activation energy of 16.8 kcal·mol^{-1}.[227,252] It is important to note that this analysis was performed in the context of a particular FAS mutant that mimicked the acyl-KS intermediate that results from STAGE 1. Though several theories have been proposed to explain the decarboxylation mechanism of the condensing enzymes involved in the fatty acid synthesis, the chemical mechanism behind this reaction is still not fully clarified and constitutes the main aspect of controversy regarding the overall condensing reaction.

Mutagenesis studies revealed that three basic residues play an essential role in facilitating the KS decarboxylation stage: the two histidines of the catalytic triad, His293 and His331, and the conserved Lys326 (Figure 6C).[32,35,252,253,277] The replacement of these His or Lys residues by Ala was found to reduce the decarboxylase activity of the KS domain by more than 2 or 3 orders of magnitude.[252] Energy-wise, the mutation of either one of the His is reflected by an increase of the Gibbs activation energy of the decarboxylation reaction by 2−3 kcal·mol^{-1}, emphasizing their importance for the decarboxylation event. Based on these experimental findings and on further structural and biochemical analyses of condensing enzymes from distinct FAS, three different mechanisms have been hypothesized to explain the decarboxylation reaction mechanism (Figure 8, STAGE 2).

A possible pathway (Figure 8, Pathway A) was proposed by Olsen and colleagues in the context of the study of the bacterial FabB enzyme (a KS homologue). It assumes that a neutral malonyl-ACP binds to the decarboxylase subsite of the acyl-enzyme complex, positioning itself near the two conserved His residues.[243] Then, His293 would initiate the decarboxylation reaction by deprotonating the carboxylic acid of the malonyl moiety, while the neutral His331 would simultaneously stabilize the building negative charge of the thioester oxo group of the malonyl moiety through a hydrogen bond, acting as an oxyanion

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

hole. An intermediate with carbanion character would emerge upon decarboxylation of the malonyl substrate, releasing the CO_2 byproduct. The conserved Lys326 could ensure the protonation state of His293 so that it could act as a base catalyst. The positive Lys326 is thought to interact with a hydroxide ion, which in turn should establish a short hydrogen bond to the protonated N_δ of His293, leaving the histidine N_ε atom free to approach the carboxylic acid of the malonyl group.

In a subsequent computational study, Lee and Engels questioned some aspects of this decarboxylating pathway, particularly regarding the protonation of the malonyl substrate and the stabilization of the carbanion-like intermediate,[276] arguing that the pK_a values of the malonic acid (pK_{a1} = 2.83; pK_{a2} = 5.69) indicate that the malonyl moiety should be deprotonated at neutral pH, and that it is very unlikely that the His331 is able, *per se*, to provide the adequate stabilization of a carbanion intermediate.

It is also noteworthy that the presence of a highly reactive hydroxide ion in the active site usually comes with an energy penalty (essentially due to its desolvation) that may (or may not) hamper the decarboxylation reaction.

In another hypothesis (Figure 8, Pathway B), Price et al. used the crystallographic structures of the *E. coli* FabB complexed with the competitive inhibitor thiolactomycin (TLM),[278] to propose that the terminal carboxylate of the malonyl moiety occupies a position adjacent to a conserved Phe and that both His293 and His331 are hydrogen-bonded to the substrate thioester carbonyl oxygen, forming a second oxyanion hole and facilitating the decarboxylation reaction leading to the release of CO_2 and the formation of the carbanion-like intermediate. In compliance with the proposal by Lee and Engels, the malonyl moiety is deprotonated at the beginning of the decarboxylation reaction, and a proper stabilization of the emerging carbanion-like intermediate is ensured by the two conserved His residues.[276] The role of the conserved Lys326 is also compatible with this pathway, as it establishes a hydrogen bond network with a water molecule, His293, and a Glu residue (Glu354 in FabB numbering, that is positionally equivalent to Glu291 of the human KS enzyme), which could control the protonation state of His293 to ensure that it can properly interact with the substrate's thioester carbonyl group.

While studying the KS domain of the animal (rat) FAS, Witkowski et al. observed that bicarbonate (and not CO_2) was released as a product of the decarboxylation reaction.[252] This led the authors to propose an alternative mechanism (Figure 8, Pathway C), where the negative charge of the carboxylate is partially neutralized by the active-site His293 upon protonation of the malonyl moiety of the malonyl-ACP substrate. The decarboxylation reaction would initiate with a nucleophilic attack by a water molecule activated through deprotonation *in situ*, or a hydroxide ion, on the C3 of the malonyl-ACP substrate, producing a negatively charged tetrahedral transition state. Again, two strong hydrogen bond interactions from His293 and His331 would be responsible for the stabilization of this labile structure. Then, the tetrahedral structure would disassemble, yielding an enolate and a bicarbonate molecule. This proposal was corroborated and further extended in 2005 by White and co-workers in a study carried out in the bacterial (*Streptococcus pneumoniae*) FabF, also a KS homologue.[196] In line with the proposal by Price et al.,[278] they further claimed that the conserved Lys and a nearby Glu controlled the protonation of His293. Again, Lee and Engels questioned that, similarly to Pathway A, a single His residue (His331) would be unable to properly stabilize the carbanion-like intermediate.[276] On the other hand, their computational calculations predicted that the N_ε of His293 is less acidic than the N_δ atom, which suggests that the first should be protonated and the latter should not. Furthermore, these authors invoked the crystallographic structure of the *E. coli* FabB-TLM complex[278] to reiterate the N_ε protonation of the conserved histidine. Indeed, if that is the case, then the N_ε of His293 cannot activate the water molecule that initiates the decarboxylation reaction of Pathway C. Nonetheless, Lee and Engels recognized that these incongruities may be due to the fact that this pathway is only applicable for the animal KS domain and, thus, may not be completely transferable to its homologues.

At the end of STAGE 2 of the chain-elongating reaction, a reactive carbanion-like intermediate is formed, and either a bicarbonate or CO_2 molecule is released as a byproduct.

STAGE 3: Claisen C−C Condensation. The third and final stage of the KS-catalyzed reaction consists in the formation of a C−C bond between the C2 of the resulting oxyanion intermediate and the acyl moiety that is bound to the catalytic Cys (Figure 8, STAGE 3). Whereas the second stage is controversial, it is widely accepted that the Claisen condensation is initiated by a nucleophilic attack of the C2 carbon of the oxyanion intermediate (that bears a carbanion character) on the carbonyl carbon of the acyl-enzyme thioester, yielding a tetrahedral transition state.[25,232,252,253] The negative charge that builds up in the carbonyl group of the Cys161-acyl is again stabilized by the oxyanion hole formed by the backbone amides of Cys161 and Phe395, similarly to STAGE 1. Finally, the labile transition state breaks down to release the β-ketoacyl-ACP product, which is ready to proceed through the fatty acid biosynthetic cycle.

3.2.3. Role of the Animal KS Domain in the Fatty Acid Chain-Length Regulation. The regulation of the fatty acid chain length results from the competition between the extension of the fatty acid chain and the chain-terminating release reaction.[206,279] Given its implication in the C−C bond-forming reaction that elongates the substrate in each elongation cycle, the KS domain plays a major role in the determination of the chain length of the final fatty acid product, in particular through its substrate specificity.

It has been experimentally demonstrated that the catalytic efficiency of the inter-thiol acyl transfer (the stage that determines the overall specificity of the KS domain) increases as the length of the acyl substrate ranges from C_2 to C_{12}, and decreases with chain lengths beyond C_{12}.[235,248] This means that the KS domain has an intrinsic preference not only for saturated acyl-ACP intermediates but also for intermediates with large chain lengths, and that it readily accepts their transfer from the ACP to its catalytic Cys161, to be further elongated. On the other hand, the longer C_{16} and C_{18} intermediates are poorly recognized by KS and, thus, less likely to be extended further. The aforementioned selectivity of this condensing enzyme is believed to be associated with the structural properties of the KS substrate-binding channel.[108,184,206] Through the analysis of the crystallographic structure of the animal (pig) FAS, Maier et al. reported the existence of a narrow constriction in the KS substrate-binding tunnel that prevents the entry of large substrates.[108] Within this structural constriction, the conserved Phe395 residue acts as a gatekeeper that regulates the access of the acyl-ACP intermediates to the catalytic Cys161 by changing the conformation of its bulky side chain.[248] The steric obstruction that is experienced in the KS binding tunnel limits

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

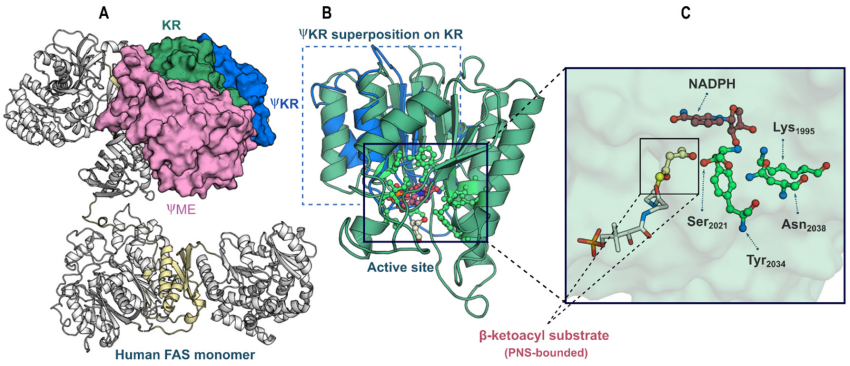


Figure 9. The KR domain and the ΨKR and ΨME domains of human FAS.[118] (A) Three-dimensional surface view of the three domains within the human FAS monomer. (B) Cartoon representation of a superposition of the ΨKR (blue) and KR (green) domains, where the similarity between both can be understood; the KR active-site region in delimited by a black line. The KR is based on the crystallographic structure of the human KR domain (PDB ID: 4PIV). (C) Close-up of the active-site cavity. The catalytic residues and NADPH cofactor are depicted in ball-and-stick representation as well as the β-ketoacyl substrate (or polyketide in KR-PKS). The PNS is bonded to the substrate but is not shown for simplicity.

the elongation of large-chain-length fatty acids such as the C_{16} palmitic acid, which remain bound to the ACP domain and, consequently, become more susceptible to the chain-termination release reaction.[206] Interestingly, in several condensing enzymes of the closely related PKS systems, the residues that are positionally equivalent to the gatekeeper Phe395 and to the other residues that form the structural constriction are variable and usually smaller, which makes the binding tunnel wider and more permissive, and allows the condensation of a larger set of acyl substrates.[108,248,280]

To conclude, even though the TE domain is the main enzyme responsible for C_{16} product fidelity, the KS domain also plays an important role in the control of the chain length of fatty acids by providing a restricted space for the growing acyl-ACP intermediate, in a way that the elongation reaction becomes compromised when the acyl chain exceeds this limited space.

3.3. Reaction Cycle of the KR Domain

After the condensation step catalyzed by the KS domain of FAS, the β-ketoacyl intermediate is modified through a sequential processing called β-carbon processing. The enzymes responsible for this process are the NADPH-dependent KR and DH and the NAPDH-dependent ER domains, the latter producing the fully saturated acyl intermediates. The β-carbon processing cycle runs seven times in total, once in each elongation turnover.

3.3.1. Biochemical Characterization of Animal KR. The KR is a keto reductase characterized by a core Rossmann fold.[281,282] KR reduces the β-keto intermediates brought by the ACP domain. It is similar to the ER enzymes/domains and to the short-chain dehydrogenase/reductase family of enzymes.[283−287] Despite this, particular aspects distinguish KR, like primary sequence identity,[257] substrate specificity, active-site architecture, and the length of an important α-helix at the NADPH-binding site.[109,288−290] The crystal structure revealed by Maier and co-workers in 2008,[108] for the animal FAS, found two additional non-enzymatic domains, both located at the periphery of KR, which were named as pseudo-keto reductase (ψKR) and pseudo-methyltransferase (ψME). In 2014 Hardwicke et al.,[169] crystallized these domains in isolated form with KR in order to identify the role of ψKR and ψME pseudo-domains. Together, these studies suggest that the ψKR corresponds to a truncated KR domain (it has half the size of the KR domain) without any catalytic capacity (Figure 9A,B). The truncated part cannot bind the cofactor NADPH, which is fundamental for the reaction. Additionally, the catalytic residues proposed to participate in the reduction reaction are non-existent in ψKR.

A ψME pseudo-domain can be found near the ψKR domain. It is homologous to enzymes of the methyltransferase family, particularly to the C-methyltransferase of the fungal PKS.[281−283] Nonetheless, experimental evidence indicates that ψME does not present any catalytic capacity as well.[284] It was proposed that the function of the pseudo-domains is structural, since the three domains—ψME, ψKR, and KR—form a compact globular structure (Figure 9A).[106,286,288,291,292] This hypothesis was supported by Hardwicke and co-workers,[169] who were unable to obtain crystals of the isolated DH and KR domains diffracting at adequate resolution, and observed that the inclusion of the ψ domains improved the crystallization of the KR domain by establishing direct contacts covering a vast area in KR.[169] Maier and co-workers[47] also suggested that the position the ψME domain facing the reaction cleft is key in helping to direct ACP access, and thus may serve to limit the diffusion space available for ACP and improve the efficiency of substrate transfer.

3.3.2. Active-Site Structure and Catalytic Mechanism of KR. A detailed analysis of the active sites of animal (human) KR, fungal (yeast) KR, and bacterial (*Saccharopolyspora erythraea*) KR-homologues FabG and EryKR1 domains/enzymes shows a remarkable structural conservation among the substrate-binding sites.[57,196] The fundamental catalytic residues include the Ser2021, Tyr2034, and Lys1995 triad (human KR numbering unless otherwise stated), plus Gln2031 and Ile2063, which interact with the β-ketoacyl substrate (KAC, with K for keto and AC for acyl) and the nicotinamide ring of the NADPH cofactor, all of them located at the bottom of the active site. In addition to this, the residue Asn2038, located in the vicinity of the catalytic residues, is also critical for KR activity (Figure 9C).[293,294] All of these residues are highly conserved and invariant between type I and type II FAS, with just the Lys1995 and Asn2038 being interchanged in bacterial (*M. tuberculosis*)

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

FabG. The same residues are conserved for the active site of the epimerase reaction.

The catalytic triad is composed by Lys1995, Ser2021, and Tyr2034. The first two were individually mutated in animal (human) KR.[295] The catalytic activity was essentially retained in the Lys1995 mutant, with the first-order rate constant decreasing by 204 min^{-1}. In the case of the Ser2021 mutant, the enzyme was inactivated. Tyr2034 was not mutated in animal FAS. However, it was mutated by an Ala in the homologous and very similar PKS enzyme EryKR1 (details in section 5),[40,109,257,271,288,290,293,296] reducing the rate constant by about 4 orders of magnitude and thus abolishing the keto-reductase activity.[295,297,298] This indicates that Tyr2034 is probably very relevant for the mechanism of the KS domain of animal FAS.

Collectively, the literature suggests stages of the KR mechanism, namely a nucleophilic attack of a NADPH-generated hydride to the β-carbon (C_β) of the KAC substrate, followed by a protonation of the alkoxide formed upon the nucleophilic attack.[109,293,299,300] The complete mechanism may be hypothesized as a nucleophilic attack of the NADPH hydride on the β-carbon, synchronously or asynchronously with the protonation of the alkoxide group (Figure 10A,B). Even though both Tyr2034 and Ser2021 are well-positioned to protonate the alkoxide (Figure 9C), Tyr2034 might restore its protonation state more easily through a proton transfer from Lys1995, probably bridge by the ribose-NADPH ring,[293] rendering the initial protonation state of the domain/enzyme. The success of the hydride transfer from the nicotinamide-NADPH ring to the substrate β-carbon (C_β, see Figure 10A) may be promoted by the good positioning of the NADPH ring over C_β[25,108,169,290,293] and the proper orientation of the Ser2021/Tyr2034 side-chain hydroxyl groups, which establish hydrogen bonds with the carbonyl group of the ACP-bound fatty-acyl intermediate that might form an oxyanion hole[301−304] stabilizing the alkoxide that is formed at the transition state of the hydride-transfer reaction.

High-level QM/MM computational[305−307] simulations were performed to elucidate this hypothetical mechanism.[293] The simulations indicated that the transition state is governed by the hydride transfer, which is in total agreement with what has been observed before in reductase NADPH-dependent enzymes.[293,308,309] The formed alkoxide anion was indeed strongly stabilized by the oxyanion hole formed by the side-chain hydroxyl groups of Ser2021 and Tyr2034 (Figure 10B), whose hydrogen bonds became shorter and stronger as the reaction path came closer to the transition state. Tyr2034 protonated the alkoxide through its side-chain hydroxide group, in a late, concerted but asynchronous transfer, forming the 3R-configured[310,311] β-hydroxy acyl intermediate (Figure 10B) at the end of the first reaction step.

The following chemical step restored the protonation state of Tyr2034, through a proton transfer from Lys1995, which was mediated by the well-positioned 3′-hydroxyl group of $NADP^+$ (Figure 10B).[293,310]

The same type of mechanism is observed in the similar epimerase reaction of the homologous EryKR1 enzyme from module 1 of the DEBS PKS.[297] The proposed mechanism for EryKR1 employs a water molecule to deprotonate the C_α atom and a Tyr to protonate the alkoxide.[294] As the tyrosine side chain is freer to rearrange in EryKR1 than in the human KR, it can perform its role without the need of a bridging group. The relatively high acidity of the EryKR1 Tyr2034 side chain (whose pK_a in water is 9.7) is attributed to an interaction with the neighboring, active-site Lys1995, that lowers its pK_a by destabilizing the anionic state.

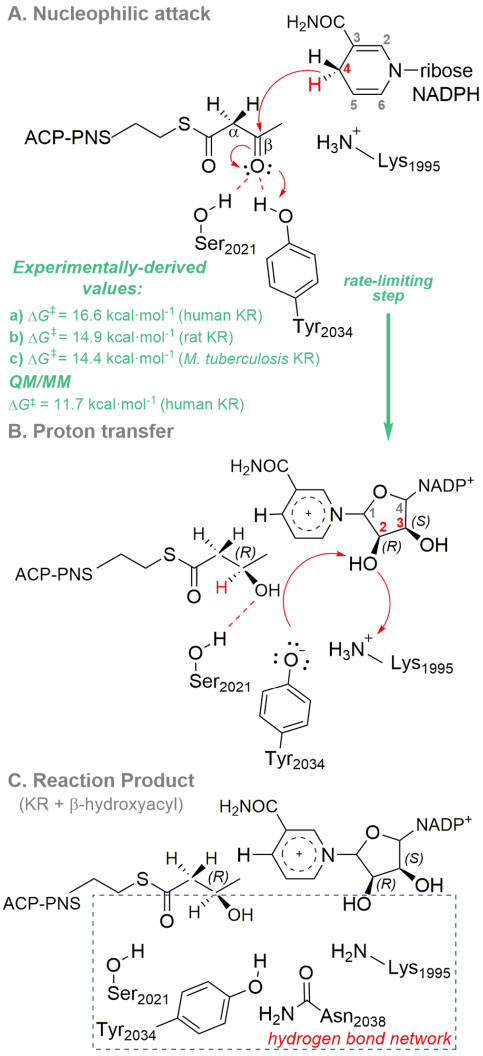


Figure 10. Catalytic mechanism for the NADPH-dependent FAS KR domain. This mechanism proceeds through two major steps: (top, A) nucleophilic attack from the hydride transfer to β-carbon of the substrate concomitantly with an alkoxide formation, (middle, B) proton transfer from ribose ring to Tyr2034-alkoxide, and (bottom, C) final product formation (β-hydroxyacyl). The Asn2038 (in C) has been included to represent the hydrogen bond network, between Ser2021, product, Tyr2034, ribose ring, Lys1995, and Asn2038. The Gibbs activation energies of 16.6 kcal·mol^{-1} and 14.9−14.4 kcal·mol^{-1} were derived from the k_{cat} values of 4 s^{-1} (25 °C) and 73−169.6 s^{-1} (25 °C), respectively,[257,288,295] using TST.

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

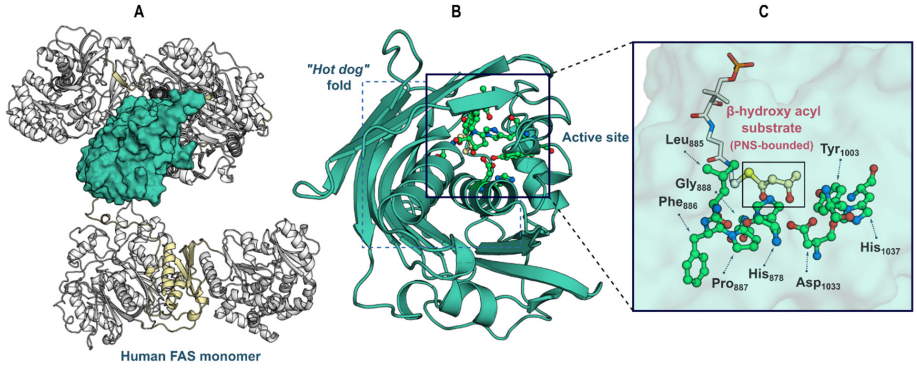


Figure 11. Dehydratase (DH) domain of the human FAS.[118] (A) Three-dimensional surface view of the DH domain within the human FAS monomer. (B) Cartoon representation highlighting the DH secondary structure elements. The active-site region (inside black lines) is embedded in the hot dog fold (blue dashed square). This DH:substrate complex corresponds to a model based on the crystallographic structure of the *Sus scrofa* DH domain (PDB ID: 2VZ9). (C) Close-up of the active-site cavity. The catalytic residues and the PNS-bonded β-hydroxyacyl substrate are depicted in ball-and-stick representation.

To conclude, participation of theTyr2034 as an acid is consistent with what is proposed for other KR enzymes in the PKS multi-enzymes, which are closely related to FAS.[108,294]

The nucleophilic attack has been referred to be the rate-limiting step of the reaction with a turnover of 4 s^{-1} in the human FAS KR domain,[295] corresponding to a Gibbs energy barrier of 16.6 kcal·mol^{-1} as derived from TST.[312] This experimental barrier is in agreement with the computational result of 11.7 kcal·mol^{-1} for the same organism.[293] Another kinetic study using animal (rat) FAS resulted in a k_{cat} of 73 s^{-1} (Gibbs activation energy of 14.9 kcal·mol^{-1}).[257] The activity of the KR domain of PKS from *M. tuberculosis* is similar to the one of animal FAS, with a rate constant reported at 169.9 s^{-1}, corresponding to a Gibbs activation energy of 14.4 kcal·mol^{-1}.[288] Overall, these results match very well with the Gibbs energy barriers determined by the QM/MM simulations.[293]

3.4. Reaction Cycle of the DH Domain

3.4.1. Biochemical and Structural Characterization. The DH domain is characterized by an $\alpha+\beta$ mixed fold in the core, with five α-helices and six antiparallel β-sheets[39] in a so-called "hotdog fold" (Figure 11).[313,314] It is noteworthy that the hotdog fold is not exclusive of the DH enzymes/domains. The acyl-CoA TE subfamily (but not the FAS TE domain) exhibits a similar folding,[70,313,315] and shares similar active-site residues; however, the acyl-CoA TEs catalyze the hydrolysis of acyl-CoA thioester substrates.

The DH enzymes/domains are found in a broad range of eukaryote and prokaryote organisms. Even though their sequences are not always very similar, they share the same catalytic residues and, probably, the same catalytic mechanism.[257,316] Structural studies revealed that the animal DH domain resembles the bacterial homologue enzymes FabA and FabZ, which are homodimers with active-site residues from both contributing monomers. The animal DH is similar to the central region of FabA/Z enzymes;[108,313,314,317] its active site is conserved.[316−319] It can also be considered as a pseudo-dimeric domain.[39,108,317,320]

3.4.2. Active-Site Structure and Catalytic Mechanism. The DH domain active site comprises an acid/base pair, Asp1033 and His878 (Figure 11C), which is highly conserved across species[39,317] (the residue numbering of the human domain will be used unless otherwise stated). This acid/base dyad is located in the C-terminus and in the N-terminal β-sheet of the same multi-enzyme domain (Figure 11B).[108,271,295,317] However, in the FabA/Z homologues, each of the two catalytic residues, belong to one of the two monomers that constitute the homodimer. Aside from the acid/base pair, an additional residue, His1037, and a water molecule may also be important for the catalysis.[39,313,317]

Studies by Joshi et al.[316] and Pasta et al.[317] reported that either the His878Ala or the Asp1033Ala mutation results in an animal DH domain unable to participate in fatty acid synthesis. Moreover, the mutation of a third residue, Gln1037 (which is replaced by a His in the human DH domain), located in the Asp1033 proximity, by an alanine, did not abolish the activity, but increased the Gibbs barrier of the rate-limiting step by ca. 1 kcal·mol^{-1},[317] suggesting a role for this residue. Taking this into account, the proposal of a catalytic triad for the human DH domain made by His878-Asp1033-His1037 is the most widely accepted in the literature.[317,318,320]

The DH catalytic mechanism was initially proposed for the FabA enzyme.[316] The mechanism for the DH domains/enzymes of other organisms have also been reported, such as those of the bacterial (*E. coli*) FabZ and the animal (*Sus scrofa*) DH domain.[39,317,318,320−323] There are two general possibilities for the dehydration: (1) a one-base mechanism (catalytic histidine acts as acid and base)[320,322] or (2) proceeding in two main steps where the acid/base role is played by two different residues. The latter option can be divided into (i) the deprotonation of the hydroxyl-acyl intermediate Cα followed by (ii) protonation of the β-hydroxyl oxygen atom to release a water molecule (Figure 12A,B). Computational studies of the catalytic mechanism of animal (human) DH domain were carried out, and the results provided solid support for the earlier proposals.[316,317,321,323] The calculations confirmed that His878 deprotonated C_α and Asp1033 protonated the β-hydroxyl oxygen, resulting in the release of a water molecule and the formation of a $C_\alpha-C_\beta$ double bond.[320] Besides the active-site triad, a further residue, Gly888 (Figure 11C), was also found to be important for the reaction,

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

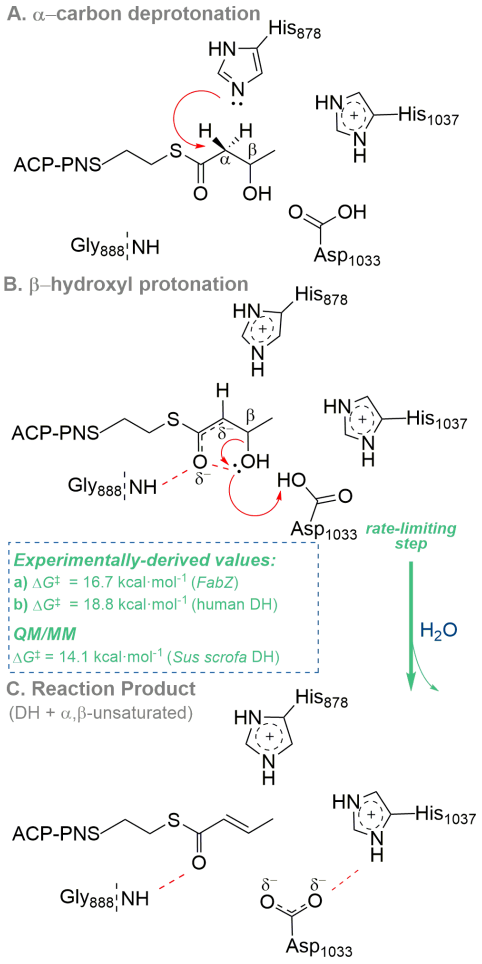


Figure 12. Catalytic mechanism of the FAS DH domain. This mechanism proceeds through two major steps: (top, A) α-carbon deprotonation by His878 and (middle, B) β-hydroxyl protonation by Asp1033, which represents the rate-limiting step of the reaction. The Gibbs activation energies of 16.7 kcal·mol^{-1} and 18.8 kcal·mol^{-1} were derived from the k_{cat} values of 3.4 s^{-1} (25 °C) and 10^{-1} min^{-1} (25 °C), respectively.[295,317] (Bottom, C) The final α,β-unsaturated product.

taking part in an oxyanion hole that stabilizes the first transition state of the reaction (Figure 12A).[39,320,321]

Deprotonation of the Substrate α-Carbon. The reaction started with the intermediate bonded to the PNS prosthetic group of ACP.[25,108,324] At the transition state of the first reaction step, His878 deprotonated the substrate Cα atom. Upon deprotonation, an enolate was formed, whose oxyanion was stabilized by a hydrogen bond established with the backbone amide nitrogen of Gly888, that became shorter and stronger as the reaction progressed toward the transition state.[320,325,326] The simulations have shown that a significant lowering of the activation Gibbs energy (9 kcal·mol^{-1}) was provided by this interaction, showing that the oxyanion hole is a fundamental part of the required interactions for the α-deprotonation reaction to occur.[320]

Competition between Asp1033 and His878 for Protonation of the β-Hydroxyl Group. Two residues, His878 (that is positively charged) and Asp1033, were in a good position to subsequently protonate the β-hydroxyl group of the resulting enolate intermediate. The more acidic residue among the two of them depends on the protonation state of a third residue nearby, His1037, which can be either positive or neutral. When His1037 is positive, Asp1033 is more acidic than His878, and when His1037 is neutral the opposite is true. Even though the computer simulation has favored the first over the second, the difference in Gibbs energy for the two alternative mechanisms is small, and a definitive picture is still to emerge.[320]

His878 and Asp1033 are both hydrogen-bonded to the hydroxyl group of the enolate intermediate. His878 stabilizes the negative charge that builds up during the elimination of the β-hydroxyl group, which resembles mostly a hydroxide ion than a water molecule at the transition state, as the Asp1033 proton is still far from the β-hydroxyl. Finally, an α,β-unsaturated product (the final product of the reaction cycle of the DH domain) is formed upon full β-hydroxyl protonation, and a water molecule is released.

The first-order reaction rate for the animal (human) DH domain is 10^{-1} s^{-1}, which corresponds to a Gibbs activation energy of 18.8 kcal·mol^{-1}.[295] The first-order reaction rate of the homologous bacterial (*E. coli*) FabZ is 3.4 s^{-1},[317] which corresponds to a Gibbs activation energy of 16.7 kcal·mol^{-1} (by using the TST).[227] The Gibbs activation energy for the rate-limiting step obtained from the computer simulations for the animal (human) DH domain was 14.1 kcal·mol^{-1}, in agreement with experimental data; the rate-limiting step was observed to be the β-elimination step, with an activation energy of ∼14 kcal·mol^{-1} (Figure 12B).[320]

3.5. Reaction Cycle of the ER Domain

3.5.1. Biochemical and Structural Characterization. The ER catalyzes the last reaction of the fatty acid elongation cycle. The chemical reaction requests a NADPH cofactor to reduce the C_α–C_β double bond to a single bond. It is supposed that the mechanism of reaction bears considerable similarities with the one of the KR domain.[293,308,309,327] The structure of the animal ER domain contains two subdomains characterized by a Rossmann-fold, which harbors the nucleotide- and the substrate-binding cavity[108,282,328,329] (Figure 13), and its overall fold belongs to the medium-chain dehydrogenase/reductase (MDR) superfamily fold.[25,108,287,330−332] This superfamily includes equivalent domains/enzymes in PKSs, quinone oxidoreductases,[333,334] Zn(II)-dependent alcohol dehydrogenases,[335−337] and crotonyl-CoA carboxylases,[338,339] among others.

3.5.2. Active-Site Structure and Catalytic Mechanism. The key residues for the catalysis are Lys1771 and Asp1797, which play an electrophilic role during the reduction reaction in domains/enzymes such as animal ER domain and the bacterial homologous enzyme and FabI.[108,340] A tyrosine or a serine plays an equivalent catalytic role in other NADPH-dependent members of the MDR superfamily, such as in the two bacterial (*Methylobacterium extorquens* and *E. coli*) Ccr and YhdH enzymes and in the fungal (*Candida tropicalis*) Etr1p enzyme.[341,342] Two additional residues, Leu1796 and

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

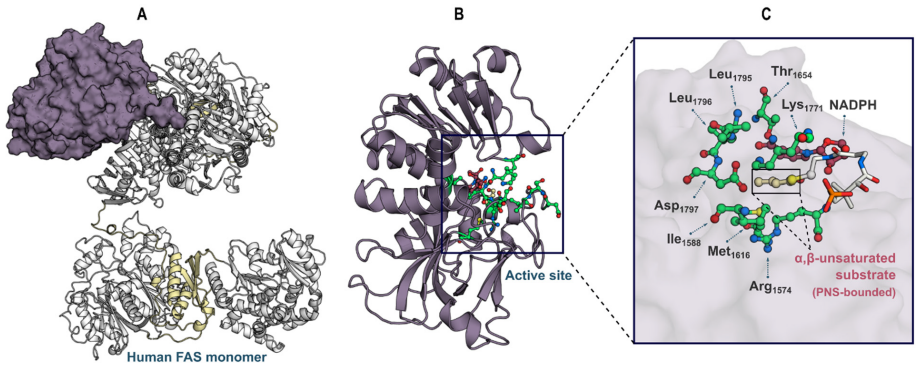


Figure 13. ER domain of the animal (human) FAS.[118] (A) Three-dimensional surface view of the ER domain within the human FAS monomer. (B) ER secondary structure elements and active-site region (within the black lines). The ER-substrate complex is taken from a computational model based on the crystallographic structure of the *Sus scrofa* ER domain (PDB ID: 2VZ9). (C) Close-up of the active-site cavity. The catalytic residues, the NADPH cofactor, and the PNS-bound α,β-unsaturated substrate are depicted in ball-and-stick representation.

Thr1654, have also been described as being important for the catalysis, with the latter highly conserved among NADPH-dependent enzymes.[343,344] The proposed role for the Thr1654 residue is to raise the barrier for an undesirable side reaction that leads to the formation of a highly toxic C_4/C_α adduct. This aspect is discussed in detail below. Raising the barrier of undesirable reactions by destabilizing the respective transition states is referred to as "negative catalysis", and it takes place together with transition-state stabilization toward the desired product exerted by the catalytic residues.[344,345]

The substrate for the first cycle of the ER domain is a crotonyl-ACP molecule bonded to the ACP through the cofactor PNS;[108,346] the catalytic mechanism of ER starts with a Michael addition of a hydride transfer from NADPH to the unsaturated C_β atom of the substrate (Figure 13C).[347,348] In a subsequent step, a proton transfer from an acidic residue (Lys1771/Asp1797) to the carbanion-like C_α would result in the saturated product (Figure 14A–C).[295,310,347−350] This mechanism has been named as "classical mechanism". However, this simple, paradigmatic, and long-lasting proposal for the double-bond reduction mechanism has been indirectly challenged by a recent work of Rosenthal and co-workers in the similar enzyme crotonyl-CoA carboxylase, which reported the experimental detection of a covalently bound adduct intermediate along the catalytic cycle.[341] Even though there were no reports of the detection of an equivalent adduct in ER (which has not been sought after anyway), this observation raised questions about the validity of the simple hydride/proton-transfer mechanism, as both enzymes/domains are believed to share a similar catalytic mechanism.

The "covalent adduct mechanism" in animal ER would start with a hydride transfer from NADPH to the substrate C_β, producing an intermediate where the C_α has a partial carbanion character. In the same or in a subsequent step, the C_α atom would attack the NADPH nicotinamide ring C_2, forming a (C_2/C_α) covalent adduct (Figure 14D).[341]

The existence of such covalent intermediate would be consistent with previous studies that have shown that an adduct was a feasible alternative in a similar solution reduction reaction, employing a NADPH-like cofactor, a number of different organic solvents, and a series of α,β-unsaturated compounds mimicking the cofactor-substrate complexed.[351,352]

A second adduct, resulting from the attack of C_α to NADPH C_4 (C_4/C_α) was also detected by Rosenthal and co-workers, using Etr1p bearing the Thr1654Val mutation (human ER numbering). The emergence of the C_4/C_α adduct, which is very cytotoxic, was precluded in the wild-type enzyme precisely by the Thr1654 residue (Figure 14), which is even more conserved than the active-site residues.[342,344] It was in this complex context that QM/MM computer simulations were carried out to try to shed some light in the controversy surrounding the true chemical mechanism of animal ER.[349] The detailed mechanistic picture that emerged is described below.

Hydride Transfer from the Nicotinamide Ring of the Cofactor or/and the Adduct Formation. Both the "classical mechanism" and the "covalent adduct mechanism" start with the transfer of a hydride from the NADPH C_4 to the substrate C_β.[349] QM/MM computer simulations have shown that the hydride transfer occurs with a Gibbs energy barrier of 14.7 kcal·mol^{-1}, leading to an enolate intermediate (Figure 14B).

It was proposed that the transfer of the hydride could be concerted with the attack of C_α (that bears a carbanion character at the transition state) to C_2 of the nicotinamide ring, resulting in the C_2/C_α adduct. The computational results show that the concerted hypothesis is not possible (a barrier of ca. 50 kcal·mol^{-1} is observed). However, after the formation of the enolate intermediate, the attack of C_α to C_2 is feasible, as the Gibbs energy barrier to overcome is limited to 17 kcal·mol^{-1} and the difference in Gibbs energy between the enolate and the adduct is −4.9 kcal·mol^{-1},[349] consistent with the experimental observation of the C_2/C_α adduct in crotonyl-CoA carboxylase.[341] The formation of the C_4/C_α adduct was also tested. It was unreachable when the attack of C_α to C_4 was concerted with the hydride transfer (Gibbs energy barrier above ∼70 kcal·mol^{-1}), and too slow when it was attempted from the enolate intermediate (Gibbs energy barrier above 27.2 kcal·mol^{-1}). When the same calculations were done in the Thr1654Val mutant, the Gibbs energy barrier was lowered to 19.2 kcal·mol^{-1}), which is kinetically feasible. This is consistent with the lack of detection of the C_4/C_α adduct in the similar crotonyl-

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

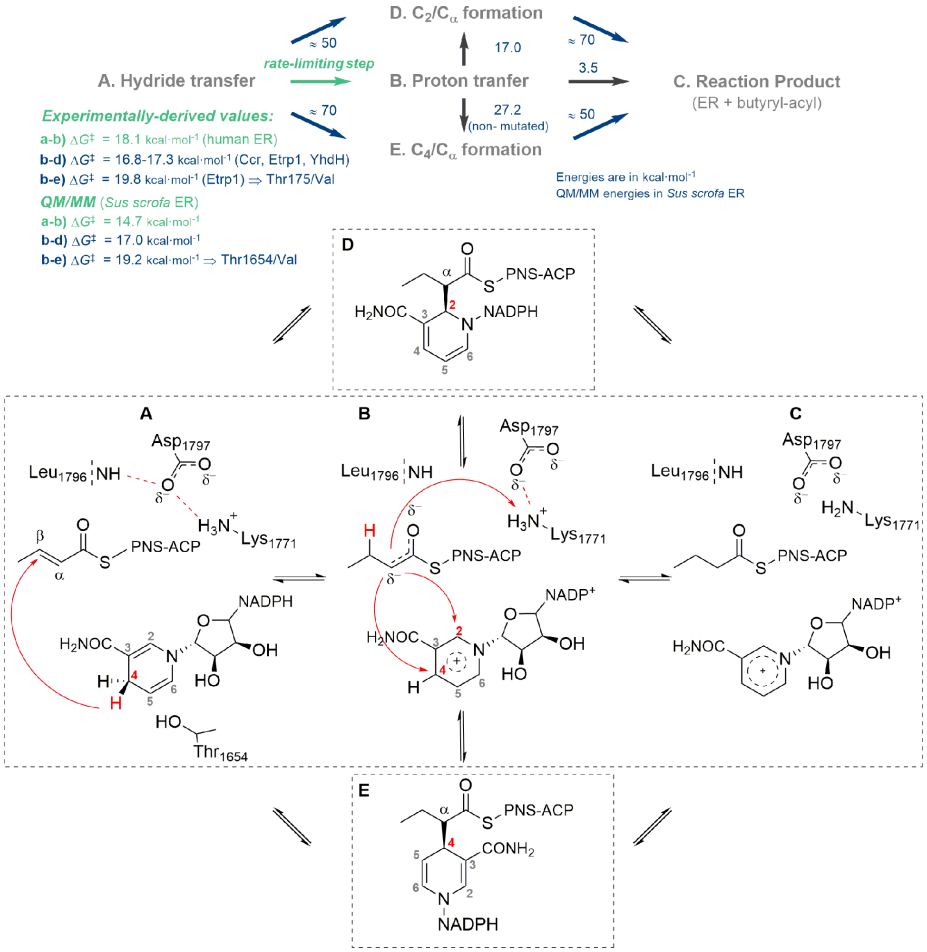


Figure 14. Catalytic mechanism for NADPH-dependent FAS ER domain. Top: Activation energies for the several possibilities for the ER reaction mechanism. The blue arrows represent reactions that are kinetically unfavorable. The Gibbs activation energies of 18.1 kcal·mol^{-1}, 16.8−17.3 kcal·mol^{-1}, and 19.8 kcal·mol^{-1} were derived from the experimental k_{cat} values of 0.33 s^{-1}, 3.2−1.3 s^{-1}, and 0.02 s^{-1} (25 °C).[295,341,342] Bottom: (A) hydride transfer from the nicotinamide ring to the β-carbon of the substrate, (B) proton transfer from Lys1771 to the α-carbon of the substrate (the salt bridge interaction between Asp1797 and Lys1771 is accentuated with a red dashed line), and (C) final product. (D) and (E) represent the C_2/C_α and C_4/C_α adducts.

CoA carboxylase wild-type enzyme and with the detection of the same adduct in the Thr1654Val mutant.

Proton Transfer from an Acidic Residue to the Substrate Cα. The next step of both the classical mechanism and the covalent adduct mechanism is the same: the protonation of C_α by an acidic residue, in this case Lys1771. It has been proposed that Asp1797 might also protonate C_α. However, its position in the animal (human) ER is far from C_α (ca. 5 Å) for an efficient protonation;[349] despite this, Asp1797 plays a critical role in the reaction, by making a salt bridge with Lys1771 (Figure 14B).[353]

As the enolate C_α bears considerable negative charge (Figure 14B), its protonation is straightforward, leading to the final product with a small Gibbs energy barrier of 3.5 kcal·mol^{-1}. Therefore, the classical mechanism is perfectly feasible from a kinetic point of view.

The Gibbs energy barriers for the protonation of either the C_2/C_α or the C_4/C_α adduct, concomitantly with the adduct breakdown to generate the product (i.e., the breaking of the C_2-C_α bond or the C_4-C_α bond in the Thr1654Val mutant), are instead very high, ∼50−70 kcal·mol^{-1}. There is thus no way to progress from the adducts to the product, and the only viable chemical pathway is the collapse of the adducts into the enolate intermediate again, from which the classical pathway can follow. The covalent adduct pathway is thus inviable. This hypothesis

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

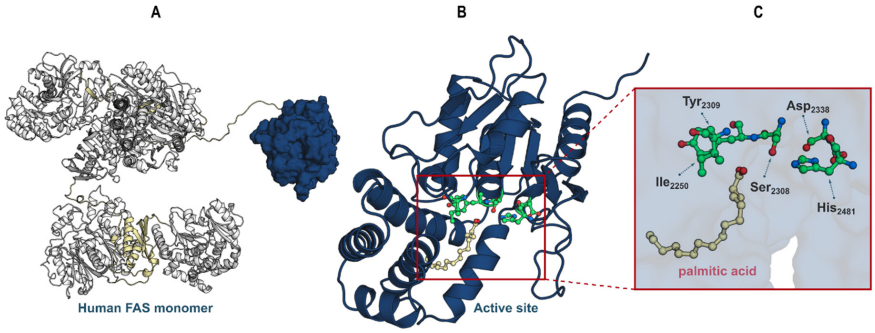


Figure 15. TE domain of the animal (human) FAS.[118] (A) Three-dimensional surface view of the TE domain within the human FAS monomer. (B) TE secondary structure elements and active-site region (within the red outlined box). The TE−substrate complex is taken from a computational model based on the crystallographic structure of the *Homo sapiens* TE domain (PDB ID: 2PX6) with a modeled palmitic acid molecule. (C) Close-up of the active-site cavity. The catalytic residues and the product are depicted in ball-and-stick representation.

was also anticipated in the past.[349] This result emphasizes that the experimental observation of a reaction intermediate does not necessarily mean that the intermediate belongs to the chemical pathway of the reaction. It can be a dead-end in the reaction landscape, as are the two adducts experimentally detected in the crotonyl-CoA carboxylase enzyme.

Several *in silico* and *in vitro* studies suggest that the rate-limiting step for the reaction cycle of the ER domain in the animal (*Sus scrofa*) FAS is the hydride transfer. The Gibbs activation energy determined computationally for this step is 14.0 kcal·mol^{-1}.[349] The experimental k_{cat} for the human ER domain is 0.33 s^{-1} (18.1 kcal·mol^{-1}).[295] The agreement between the *in silico* experiments and the *in vitro* assays for the animal ER is good.

The k_{cat} for $C_2/C\alpha$ adduct formation in the Ccr, YhdH, and Ert1p enzymes is in the range of 3.2−1.3 s^{-1}, which correspond to a Gibbs activation energy of 16.8−17.3 kcal·mol^{-1}, emphasizing the good agreement between experiments and computation (17.0 kcal·mol^{-1}). The rate constant for the C_4/C_α adduct formation is k_{cat} = 0.02 s^{-1} in the mutant Etr1p enzyme,[341,342] which correspond to a Gibbs activation energy of 19.8 kcal·mol^{-1}, comparing very well with the computational value of 19.2 kcal·mol^{-1} obtained for the animal (*Sus scrofa*) ER mutant.[349]

3.6. Reaction Cycle of the TE Domain

3.6.1. The TE Superfamily. The TEs constitute a large superfamily of enzymes present in all living cells. Within its 25 families, and a recently proposed 26th one,[354] TEs catalyze the cleavage of thioester bonds on a wide array of cellular substrates: acyl-CoA, acyl-ACP, protein substrates, and acyl-glutathione. The TE of megasynthases, such as FAS, PKS, and non-ribosomal peptide synthetases (NRPS), belong to the families TE16−TE19, which are characterized by a very conserved α/β hydrolase fold,[355] that specifically catalyze either cyclization, termination, or the removal of defective products by hydrolysis.[356−358]

3.6.2. Animal FAS TE1 and TE2. There are two distinct TEs involved in the *de novo* fatty acid biosynthesis. TE1 refers to the TE domain in animal FAS; it is the catalytic domain responsible for the off-loading mechanism in the fatty acid synthesis process. After seven to eight elongation cycles, TE1 catalyzes the release of the cargo, a long-chain palmitoyl (C_{16}) or stearate (C_{18}), from the PNS arm of the FAS ACP domain.[70,108]

TE1 is bound to a flexible linker stemming from the KR domains, present in both FAS polypeptide chains, and its position precedes that of the ACP domains (Figure 15A). TE1 is easily isolated from the megasynthase body through proteolysis;[359] when detached, the FAS products shift from 16 or 18 to 20 or 22 carbon-long acyl chains. This shift occurs due to the loss of TE1 action, responsible for the tight regulation of the size of its products.[9] Therefore, when fully functioning, the activity of TE1 is strongly correlated to the length of the acyl substrate, hitting its peak when the substrate is C_{16} and being lower for chains longer than C_{18} or shorter than C_{16}.[359−361]

Alongside TE1, the monofunctional, discrete type II TE2 enzyme can also interact with FAS, causing the release of shorter length fatty acids, C_{12} and C_{14}. TE2 is additionally known as oleoyl TE and medium-chain *S*-acyl FASN thioester hydrolase (E.C. 3.1.2.14; Uniprot Q9NV23). This enzyme was initially identified in the mammary gland epithelial cells of rats,[362,363] and its expression is upregulated during lactation.[364,365] Considering that human milk fat contains 28% of C_{16} and C_{18} fatty acids, and ∼45% C_8 to C_{14} lipids, TE2 plays a significant role in the synthesis of fatty acids present in human milk.[366]

3.6.3. Structure Overview. Both TE1 and TE2, members of the TE-C clan of the TE superfamily,[367] present a canonical α/β hydrolase fold with two subdomains, A and B. The α/β-hydrolase fold family was first described in 1992,[368] and consists of a β-sheet of eight strands, with the second strand antiparallel to the others and ordered as 12435678.[369] Interestingly, the structure of the TE1 FAS domain (Figure 15B) does not strictly follow this arrangement—its subdomain A displays only seven β strands, lacking the N-terminal β1 strand. In regard to the α portion of the canonical α/β hydrolase fold, it is made up of four α-helices flanking the half-barrel,[70] with a loop insertion near the catalytic triad site, between strands β5 and β6. In the context of TE1, this loop translates to an entire all-helical atypical subdomain B, crucial to the TE1's activity.[175]

On a closer look, the catalytic triad of the α/β-hydrolase fold family involves a nucleophilic serine, a Asp/Glu, and a histidine (Figure 15C). In the canonical organization, the nucleophile is positioned in the loop between strand β5 and helix α3, the Asp/Glu in the loop between strand β7 and helix α5, and finally, the histidine between strand β8 and helix α6.[217] In the case of the animal TE1 catalytic triad, both the nucleophile (Ser2308) and

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

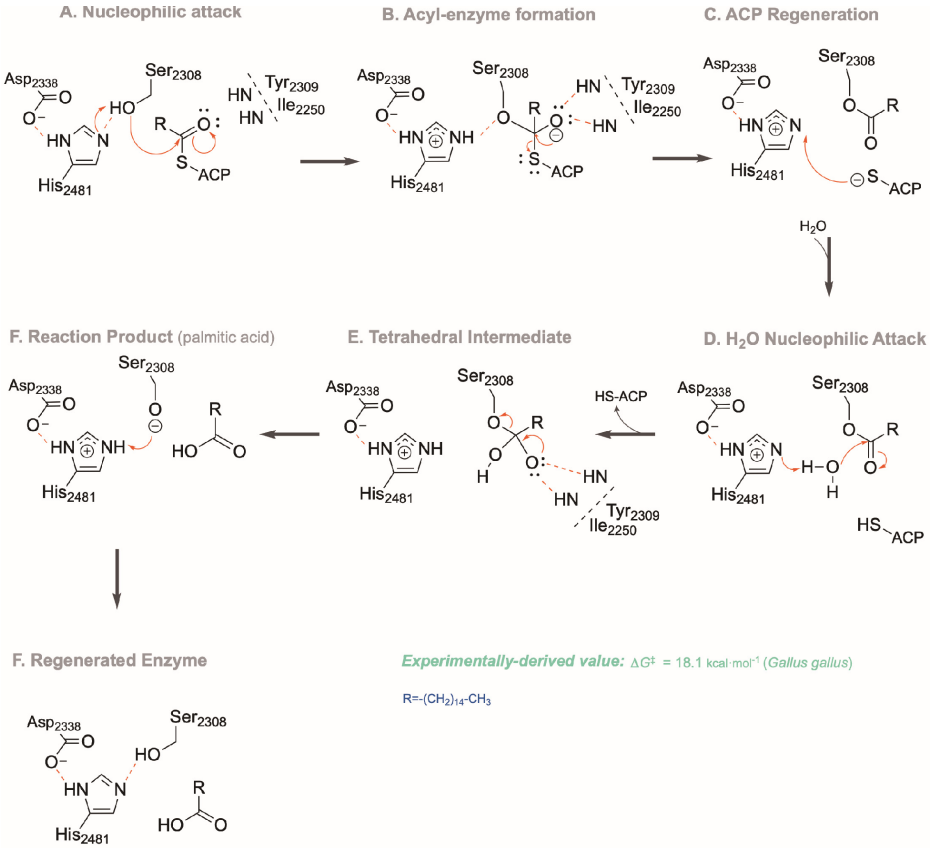


Figure 16. Catalytic mechanism proposed for FAS TE domain. Proposed mechanistic pathways for the TE domain catalytic mechanism. ("R" stands for the acyl moiety of C_{16} that will be cleaved by the TE domain. The residues are numbered according to the human TE domain.) The Gibbs activation energy of 18.1 kcal·mol^{-1} was derived from the k_{cat} value of 0.32 s^{-1} (25 °C)[382] using TST.

the His2481 follow the canonical positioning, but the Asp2338 is located in the loop connecting subdomain B β5 strand to the subdomain A helix α5. The human TE1 subdomain B is aptly positioned in relation to the catalytic triad. This positioning is unique to human FAS TE1, suggesting a role in the specificity of the enzyme.[70]

Solved structures of the TE domain are abundant, offering insight into its organization: the PDB contains TE structures belonging to the type I FAS from animals (human[70,112,175] FAS TE domains) and fungi (the PKS TE domain of *Aspergillus parasiticus*[370]), as well as TE structures derived from bacterial type II FAS (the enzymes from *Bacillus subtilis*[371] and *E. coli*[372,373] and the NRPS TE domain from *B. subtilis*[374,375]). Interestingly, the structure of fengycin TE (FenTE; PDB 2CB9) suggests the existence of a lid domain (corresponding to the subdomain B), above the active-site region, acting as a control mechanism to prevent the access of unwanted substrates.[375] This mechanism is not observed in the human FAS TE1 (more on the topic below).

3.6.4. Catalytic Mechanism. TE1's function is to carry out the final step of *de novo* fatty acid synthesis. It catalyzes the release of the C_{16} palmitic acid by hydrolysis of the acyl-*S*-PNS thioester bond at the ACP domain. Given its higher affinity to chain lengths of C_{16} or C_{18}, its activity is only observed once the fatty acid chains are released from the FAS complex. The key moments for the TE1 role in the fatty acid synthesis include (i) establishing transient protein:protein interactions between ACP and TE1, (ii) recognizing the substrate chain length, and (iii) catalyzing the cleavage of the fatty acid chain from the ACP, releasing the product into the cytosol.

3.6.4.1. Recognition of the Substrate's Chain Length. The affinity of TE toward palmitic acid depends on the size and shape of its hydrophobic binding groove. In this groove, the substrate and active-site-surrounding residues establish hydrophobic interactions, keeping the fatty acid correctly oriented toward the active site. A molecular docking assay provided insight into the fitness of binding of 7 distinct ligands in the human TE domain: the predicted binding free energy supports the idea of

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

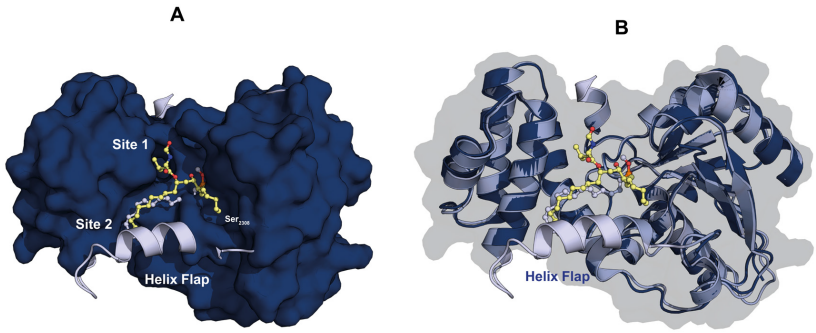


Figure 17. Structure of the binding site of the TE domain of the animal (human) FAS.[118] (A) Close-up three-dimensional surface view of the TE domain within the human FAS monomer (PDB ID: 2PX6). The active-site cavity shows two distinct binding sites: Site 1, for short-length C chains, located at the interface between the two TE subdomains; and Site 2 for the C_{16}–C_{18} substrates. The helix flap (PDB ID: 3TJM) highlighted in gray cartoon is responsible for the formation of the deep groove-tunnel Site 2 that harbors the C_{16}–C_{18} substrate. The ligand Orlistat is represented in yellow ball-and-stick representation, with its peptidyl moiety protruding into the Site 1 (resembling the behavior of short-length carbon chains. The MGLFP ligand is represented in gray ball-and-stick representation. For clarity, the active-site residues are absent, except for the indication of the position of the Ser_{2308} that covalently binds the substrates. (B) Cartoon representation of both PDB ID: 2PX6 and 3TJM, aligned, displaying the helix flap in its disordered form (blue, PDB ID: 2PX6) and ordered form (gray, PDB ID: 3TJM).

the TE optimum catalytic activity with a C_{16} substrate.[376] More on this subject in the following section 3.6.5.

3.6.4.2. Catalytic Mechanism. Despite the efforts of the scientific community, there is not a catalytic mechanism supported with experimental evidence for the TE1 domain of the animal FAS. Its active site is located at the subdomain A and exhibits a canonical catalytic triad, which encompasses residues Ser2308-His2841-Asp2338.[175] A survey of proposed mechanisms and studies for hydrolases making use of a similar catalytic triad[70,112,376−381] suggests that the cleavage of the substrate is supposed to require two distinct steps. First, the catalytic Ser2308 hydroxyl should be deprotonated by His2481, resulting in an enhanced nucleophilic character at the Oγ of the Ser. The nucleophilic Oγ then attacks the carbonyl carbon of the substrate, forming an acyl-enzyme tetrahedral intermediate, with the resulting acyl oxyanion stabilized by an oxyanion hole formed by the backbone amide nitrogens of Tyr2309 and Ile2250. The resulting nucleophilic PNS can then deprotonate the positively charged His2481, which is then available to deprotonate a water molecule to form a hydroxide nucleophile prone to attack the carbonyl carbon at the acyl-enzyme ester, resulting in a second negatively charged tetrahedral intermediate. Finally, the unstable intermediate is resolved upon deprotonation of His2481 by Ser2308, regenerating the active site and releasing the carboxylic acid product (Figure 16).

The proposed mechanism is partially based on the principle of convergent evolution; catalytic triads are some of the best examples of it. Chemical constraints on catalysis have led to the same catalytic solution independently evolving in at least 23 separate superfamilies.[383] Their mechanism of action is consequently one of the best-studied in biochemistry, and it is reasonable to assume it as the most plausible possibility for TE. Further studies are still needed to definitively validate (or deny) the present proposal.

3.6.5. Substrate Binding Sites and Substrate Specificity. During the fatty acid synthesis process, the growing fatty acid chain is attached to the PNS cofactor of the ACP domain, and it is released by the endogenous C-terminal FAS TE1. The high selectivity of the human FAS TE1 depends on its ability to interact exclusively with the C_{16} palmitoyl acyl substrate. Hence, there is a requirement for an appropriately sized substrate channel, with a mostly hydrophobic character, consistent with the nature of its substrate.

Prior to the release of the crystal structure of the TE1 domain, the TE1 substrate pocket was thought to be at the interface of the subdomains A and B, with about four to six carbons fitting into the interface cavity.[70] In 2017, the first structure of human TE1 was solved with its inhibitor Orlistat at 2.3 Å resolution (PDB ID: 2PX6)[175] and challenged this idea. The structure shows both the product state and an acyl-enzyme intermediate. Focusing in the acyl-enzyme intermediate, it is possible to observe an ester linkage between Ser2038 and the C1 carbon atom of Orlistat, consistent with the nucleophilic attack at the carbonyl position. This frozen frame provided a detailed view into the catalytic mechanism of human FAS TE1, and offered an unexpected insight into the acyl-enzyme complex. In the acyl-enzyme conformation, the electron density map shows Orlistat scaffold elements (peptidyl moiety, palmitic core, and hexanoyl tail) covering three different key regions of the TE domain in this step of the catalysis. However, this structure was missing a key element for the substrate specificity mechanism: the amphipathic α-helix, called the flap (residues 2342−2355). This α-helix residue was first observed when Zhang and co-workers crystallized the human TE1 in complex with another substrate: a C_{18} polyunsaturated fatty acid called methyl γ-linolenylfluorophosphonate (MGLFP) (PDB ID: 3TJM).[112] Comparison of the X-ray structures of human TE-Orlistat complex[175] and human TE-MGLFP complex[112] reveals that, despite being very similar, there is a stark difference: the existence of an amphipathic α-helix, called the flap (residues 2342−2355) (Figure 17B), visible in the second crystal structure but previously unsolved in the first. This flap helix establishes several complementary interactions with the γ-linolenyl chain of the substrate, creating a hydrophobic groove-tunnel that harnesses the substrate; interestingly, the flap helix is absent in the ligand-free structure determined by Zhang et al., suggesting that gatekeeper α-helix could have arisen by a γ-linolenyl chain-facilitated formation of the helix from a

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

disordered polypeptide segment or a stabilization of a very flexible α-helix.[112]

In sum, the works by Pemble et al. and Zhang et al. have shown the intricacies behind the mechanism for substrate specificity in TE. As PNS carries the substrate into the active-site region, depending on its length, it can either bind in Site 1 at the interface between the two subdomains (up to 12 C atoms) or bind at Site 2 (longer than 12 C atoms) (Figure 17A). Spanning the two TE subdomains, the Site 2 is thought to arise from the interplay between the disordered α-helix and the acyl-chain ligand which causes concomitant ordering and involvement of a helix flap. The well-defined groove-tunnel Site 2 starts near the Ser2308 in the main subdomain, and terminates deep in the tunnel in the small subdomain formed by the triangulation of the three helices: the helix flap, α5, and α8. This finding substantiates the importance of the two subdomains combined. Moreover, at the end of the Site 2, the residues Arg2428, Glu2431, and Tyr2424 (helix α8) are thought to be responsible for docking the C_{16}–C_{18} substrate into place, while generating a barrier to longer substrates.

3.6.6. Interplay between TE1 and TE2. Investigations into the structure and substrate specificity of human TE2 have allowed for comparison of and further insight into the binding site of the acyl chain, enzymatic activity, and interaction with the ACP domain. While the human TE1 is significantly more active on the acyl-CoA substrate than TE2, when these (acyl-CoA substrates) are bound to the ACP domain, the TE2 enzymatic turnover increases in ~4 fold. Additionally, when full-length human FAS is incubated with TE2, the results are a fatty acid profile enriched in short chain fatty acids.[366]

A comparison between human TE1 and TE2 has shown that, structure-wise, the subdomain B of TE2 corresponds to a capping domain. While the serine hydrolase core is preserved, this capping domain acts as a blocking device to prevent the access of the substrate to the catalytic triad, resembling the mechanism found in the TE of other megasynthases, such as the PKS TE II (SfrTE,[374] SrfTEII,[384] RifRTEII[385]) and the NRPS TE I (EntF NRPS[386] DEBS TE[387,388]). For the first, the capping domain moves, acquiring distinct conformations to regulate the access of substrates, to adapt the shape and size of the substrate chamber, and to recognize cognate ACP.[374,384−386] However, for NRPS TE I, the capping domain remains in a fixed position, crucial for the maintenance of the structure of the active-site tunnel.[387,388] The human TE2 lid domain displayed great similarity to the correspondent lid domains in the TE of type II PKS,[366] suggesting that its lid domain might also be mobile, and an adjacent coil region may rearrange to accommodate acyl-ACP docking and catalysis. Together with the information that TE2 has the ability to readily outcompete the TE1 domain of human FAS in a pool of C_{16}-ACP, but displays low activity when incubated with only acyl-CoA,[366] it suggests that TE2 is capable of selectively interact with the human ACP FAS domains. This interaction requires a shift in conformation of TE2, from closed to open, for the acyl-PNS to be loaded into the active site. These observations are useful to further investigate the workings in the acyl-ACP and TE1 interaction.

Members of the α/β hydrolases superfamily present the same chemistry, with the outstanding ability to catalyze the hydrolysis and cyclization of a wide variety of substrates. The selectivity toward specific molecules derives from the surface loop insertion into the canonical α/β core of superfamily. In the case of the human FAS TE1, with the Orlistat-TE1 complex, we are offered evidence that the surface loop insertion acts as a controlling device that provides a hydrophobic surface channel that effectively binds the right-sized substrate. One particular case where such advantage has been recognized, is the considerable interest in the microbial production of medium-chain fatty acids as alternatives to fossil fuels.[357,389,390] TE specificity and hydrolase activity allow for the production of microbial fatty acid derivatives of specific size, with examples including methyl and ethyl esters, fatty alcohols, and alkanes.[391,392]

3.7. Overall Reaction Cycle of FAS

The synthesis of a palmitic acid molecule by FAS is an intricate process that combines the synchronized action of multiple catalytic domains, which are responsible for a panoply of chemical reactions including reductions, acyl transfers, hydrolysis of thioesters, and dehydrations. Throughout the six preceding sub-sections, the catalytic phenomena concerning these specific reactions were presented and discussed in depth. The present sub-section consists of a brief overview on the energetics and the major atomistic/functional details regarding the overall reaction cycle of FAS.

3.7.1. Kinetic Rates of the Individual Reactions of Fatty Acid Synthesis. Figure 18 compiles a set of experimentally derived and a set of computationally determined Gibbs activation energies ($\Delta G^{\ddagger}$) regarding the reactions catalyzed by each animal FAS domain. While the first was derived from the k_{cat} values reported in the literature in light of the TST,[227] the latter was determined through computational QM/MM calculations. It is noticeable that the experimental $\Delta G^{\ddagger}$ energies lie close to each other and fall within a narrow interval of values (15.6−18.8 kcal·mol^{-1}), which suggests that no particular catalytic domain dictates the overall kinetics of the animal FAS. Thus, one may hypothesize that the high efficiency of the animal FAS is not related to a specific chemical step performed by any of its catalytic domains, but most likely to the peculiar organization

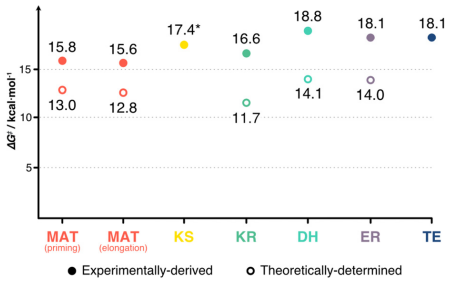


Figure 18. Gibbs activation energies associated with each catalytic domain of the animal FAS. The experimental $\Delta G^{\ddagger}$ values (filled dots) were derived from the k_{cat} data that was presented in sections 3.1−3.6, using TST and considering the original animal and temperature conditions in which each k_{cat} was determined.[224,251,295,382] The k_{cat} values of the MAT and TE domains were determined in the single domains plus PNS/ACP. The k_{cat} values for the domains KS, ER, KR, and DH were determined in the complete FAS multi-enzyme. The theoretical $\Delta G^{\ddagger}$ values (open dots) were determined employing computational QM/MM methods.[223,226,293,320,349] Note: The $\Delta G^{\ddagger}$ value for KS concerns the inter-thiol acyl-transfer stage of the animal KS-catalyzed reaction, which, among the two known stages with published k_{cat} values (the inter-thiol acyl transfer and the decarboxylation of the chain-extender malonyl-ACP), is the one associated with a higher activation Gibbs energy.

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

of these domains in a megasynthase structure that should minimize issues regarding substrate diffusion or enzyme concentration. In section 4, we will provide a detailed discussion regarding this subject.

The systematic difference between the $\Delta G^{\ddagger}$ values of the experimental and the theoretical data sets is expected and worthy of a minor comment. Given that the theoretical QM/MM calculations strictly concern the chemistry of the catalytic reaction under study, the resulting $\Delta G^{\ddagger}$ energies (open dots in Figure 18) are predicted to be lower than those derived from the experimental k_{cat} values (filled dots in Figure 18), as the latter not only contemplate the chemistry, but also the substrate's binding and the product release. This is, indeed, reflected in Figure 18. In general, computational inaccuracies lead to overestimations in the Gibbs activation energies, which is quite the opposite of the systematic trend seen in Figure 18. As such, it might be suggested that ACP-bound substrate transport and ACP binding/unbinding should have roles in the overall kinetics.

Nevertheless, by resting within a tight range of $\Delta G^{\ddagger}$ values, the theoretical energies clearly follow the trend of the experimentally derived ones and also suggest that no particular domain catalyzes a reaction whose chemistry is rate-limiting of the overall fatty acid biosynthesis.

The theoretical−experimental synchrony is particularly evident in what concerns the study of the catalytic machinery of the animal FAS. In general, the computational QM/MM results supported and validated mechanistic proposals that were formulated based on experimental evidence. Furthermore, by allowing an atomistic-level description of the chemical reactions catalyzed by the animal FAS, these theoretical methods unveiled important catalytic details that are transversal to the multiple FAS domains and that would be extremely difficult to determine experimentally.

3.7.2. Molecular View of the Multi-enzyme Reaction Mechanism. The molecular outline of the reaction mechanism of fatty acid synthesis by FAS in Figure 3 can nowadays be systematized with a much higher level of detail, as an atomic-level insight into the machinery of this megasynthase is coming to light.

Fatty acid synthesis initiates with a primer acetyl-CoA that is loaded into an ACP-PNS complex by MAT (Priming Cycle in Figure 3). The reaction proceeds through (1) the rate-limiting acetylation of the nucleophilic Ser581 of MAT that leads to the release of CoA and (2) the deacylation of Ser581 by the nucleophilic sulfur of the ACP-PNS complex, to load the ACP for further fatty acid initiation. Both stages rely on a MAT catalytic His683 to act as base, deprotonating and thus activating the nucleophiles, and result in tetrahedral intermediates that are crucially stabilized by hydrogen bonding provided by an oxyanion hole constituted by the backbone of MAT residues Met499 and Leu582.[223,226]

Upon this initiation cycle, the first of the (typically) seven elongation cycles of palmitate synthesis begins. The ACP-PNS complex covalently carries the acetyl primer to KS, where a Cys161-His293-His331 triad is present to catalyze the acetyl transfer to KS through a nucleophilic attack by Cys161. Again, hydrogen-bonding to an oxyanion hole constituted by the backbone of the catalytic Cys161 and Phe395 bridges KS and the ACP-PNS complex,[34,35,232,243,244,246,252] and the ACP-PNS complex is released soon afterward. The first loading of ACP-PNS with malonyl is again catalyzed by MAT (Elongation Cycle in Figure 3), this time using malonyl-CoA as a substrate, and the reaction mechanism proceeds in a similar fashion to that of the initiation step. The positively charged Arg606 at the active site of MAT plays an important role on its affinity toward malonyl.[104,223,223,226]

Upon loading the malonyl extender into ACP-PNS, a series of transformations by the processing KS, KR, DH, and ER domains/enzymes will take place, resulting in a saturated fatty acyl substrate ready for a new elongation cycle. The ACP-PNS loaded with malonyl will again bind to KS, where the malonyl extender now undergoes a rate-limiting decarboxylation and a Claisen condensation with the acetyl primer covalently bond to Cys161 of KS, in order to produce a covalent complex between ACP-PNS and the resulting β-ketoacyl.[251,252,268,276] While the Claisen condensation is not thought to involve significant changes in the active site of KS, the decarboxylation is thought to be a water-mediated reaction, with a conserved Lys326 acting as a base, and His293 and His331 forming an oxyanion hole to stabilize the resulting carbanion.[196,243,252,278] However, a detailed mechanism remains elusive. The resulting ACP-PNS-(β-ketoacyl) proceeds for a β-keto reduction by the NADPH-dependent KR, which occurs through NAPDH oxidation to $NADP^+$ assisted by an oxyanion hole from the catalytic Ser2021 and Tyr2034. The β-hydroxyacyl intermediate is obtained upon a simultaneous acid−base reaction involving Tyr2034 and Lys1995.[293] Dehydration of the ACP-PNS-(β-hydroxyacyl) then takes place through two acid−base reactions in which residues His878 and Asp1033 perform a direct role: His878 deprotonates the α-carbon of the β-hydroxyacyl, and Asp1033 acts preferably as an acid to carry out the rate-limiting dehydration of the β-carbon and produce ACP-PNS-crotonyl. As in MAT, an oxyanion hole with the backbone of a Gly888 and the β-hydroxyl of the substrate is formed with the thioester carbonyl of the malonyl group, and is determinant for the kinetics of α-carbon deprotonation.[320] The reduction of the resulting ACP-PNS-crotonyl product to ACP-PNS-butyryl is finally concluded in the NADPH-dependent ER. Similarly to KR, the reduction of ACP-PNS-crotonyl to the final ACP-PNS-butyryl by ER involves NADPH oxidation to $NADP^+$, followed by a fast acid−base reaction by an acid Lys1771.[349]

The final ACP-PNS-butyryl can now bond covalently to the KS to start the second of the seven elongation cycles that lead to the production of the main palmitate product, after which it binds the TE hydrolase for fatty acid release. Premature productive binding to TE and fatty acid release is possible but a rare event. Similarly to MAT, the catalysis by TE is thought to rely on a catalytic Ser2308-His2841 dyad, which in the case of TE composes the Ser2308-His2841-Asp2338 triad that is common of the serine hydrolases class of enzymes;[175] however, once the acylation of the Ser2308 is achieved, a hydrolysis reaction leads to the release of the final palmitic acid product. The recurrent oxyanion hole, here formed by the backbone of Ile2250 and Tyr2309,[175] is expected to stabilize the intermediate oxyanion characteristic of the serine hydrolase family of enzymes; however, a detailed view of the reaction mechanism of TE is still lacking.

3.7.3. Chemical Highlights of the Reaction Mechanism of FAS. While the residues catalyzing explicit chemical bond formation and cleavage in each FAS domain/enzyme were successfully identified from both X-ray crystallography and site-directed mutagenesis experiments, the recent computational studies have identified roles yet to be clarified from residues contributing both to substrate binding and catalysis. A critical aspect of fatty acid synthesis by FAS seems to be the systematic

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

use of the oxyanion hole arrangement to stabilize oxyanions generated in ketide groups,[223,226,293,320,349] which are observed across every step of fatty acid cargo, elongation, and release. Estimates from computational studies of animal FAS domains indicate that the oxyanion hole mechanism can contribute in more than 3 kcal·mol^{-1} for TS stabilization and up to 11 kcal·mol^{-1} in intermediate stabilization,[223,226,293,320,349] which renders them as a key arrangement for efficient fatty acid synthesis.

While for MAT, KS, and TE, the oxyanion hole arrangement is actively involved in the stabilization of tetrahedral intermediates formed during fatty acid cargo, condensation, and release,[223,226] during β-ketide elimination by KR, DH, and ER, the arrangement most likely prevents the formation of thermodynamically unstable carbanion intermediates by displacing electron density toward the neighboring ketide oxygens.[293,320,349] The oxyanion is always located at the thioester carbonyl of the malonyl extender (with the only exception of the β-keto reduction by KR, where the oxyanion is formed at the β-ketide of the acyl substrate), which may suggest that the thioester carbonyl contributes for the proper binding of each malonyl extender and the specificity of animal FAS toward palmitic acid synthesis. Altogether, fine-tuning the oxyanion hole arrangement may be a way to develop FAS mutants with specificity toward other acyl extenders or with broader functionalization potential.

Another aspect that can be seen throughout the FAS cycle is the abundance of bulky, and often hydrophobic, residues in the active site of each domain/enzyme, which grows from KR to ER, with an increasing count of Leu, Pro, or Phe. Such a pattern of increasingly bulky and hydrophobic active sites, which is not observed in the PKS family,[108] might contribute to FAS synthesis's specificity toward long-chain saturated fatty acids. The oxyanion hole that stabilizes the thioester carbonyl in almost all stages of fatty acid synthesis also predominantly involves the backbone of hydrophobic residues, namely Met499 and Leu582 in MAT,[223,226] Cys161 and Phe395 in KS,[243,252] Gly888 in DH,[320] or Ile2250 and Tyr2309 in TE.[175] The bulky side chain of the conserved Phe395 of KS also regulates the access of the growing substrate to the catalytic core.[206,248] In addition, the conserved residues Leu1796 and Thr1654 from ER are also pointed out as necessary for the activity of ER, and particularly Thr1654 is suggested to contribute to the specificity of ER by precluding undesirable side reactions leading to toxic side products.[344,345]

Despite these preliminary observations, a complete atomistic depiction of the role of every binding site of the domains/enzymes involved in fatty acid synthesis remains elusive, and further studies are required for a better portrayal of FAS capabilities as a synthetical nanofactory.

4. WHY HAS EVOLUTION ALLOWED THE EMERGENCE OF COMPLEX MEGASYNTHASES?

The fact that evolution has allowed the emergence of complex megasynthases remains a mystery to date. Many are the hypotheses that have been put forward, but all of them seem to have some flaw, and we shall play devil's advocate in this section, hoping that the reader might get some good ideas out of it all and enlighten the present scientific community, which remains in the dark with regard to this subject.

4.1. Animal, Fungal, and Bacterial FAS

In nature, there are many examples in which specific metabolic pathways are confined within a delimited space. For instance, in multi-cellular complexes, chemical pathways that could not co-exist are physically separated, and others that otherwise would be thermodynamically unfavorable are coupled.[393] Within the cell, the different compartments aggregate pathways that need certain cellular conditions, and that may be harmful or impair other pathways if they occurred in other compartments or the cytoplasm.[393] The peroxisome, for instance, provides a micro-environment with an alkaline pH, which enhances the catalytic activity of peroxisomal oxidase and catalase.[394,395] In addition, the peroxisome contains the toxic hydrogen peroxide to be degraded by catalase.[396]

At the protein scale, enzymes involved in the same metabolic pathway are, more often than not, close in space due to the needed protein:protein interactions between them, instead of homogeneously distributed within the cell or compartment.[397] The high protein concentration within a cell (estimated to be as high as 200−300 mg/mL or about 30% of the volume of a cell[398]) further promotes those protein−protein interactions.[399] The fact that physical proximity is an established advantage supports the idea that the formation of multi-domain and multi-enzyme complexes contributes to enhance the efficiency of metabolic pathways.[393]

The most logical explanations as to why enzymes, involved in the same pathway, benefit from being physically assembled into a multi-enzyme are (i) the so-called substrate channeling, i.e., the facilitated transfer of the product of one multi-enzyme domain active cycle to the next in the pathway, without the need of extensive aqueous-phase diffusion, which avoids the equilibration and dilution of the intermediates within the bulk solvent, and (ii) to avoid its eventual capture by a competing enzyme or chemical reaction with other solutes.[399] This unique catalytic behavior of multi-enzymes is suggested (iii) to allow catalytic properties unreachable by pathways composed of individual, non-associated enzymes,[400] as in the processing of short-lived/unstable intermediates[401] or in the interaction between enzymes that may stabilize active conformations and allow a coordinated control of enzyme activity.[402] Besides, molecular crowding within the cytoplasm (even more significant in some cellular compartments) hinders solute diffusion, benefiting compartmentalized metabolic pathways, from which multi-enzymes represent the ultimate level of sophistication.[399] Other benefits include (iv) the avoidance of harmful effects for other cell constituents in pathways in which the intermediates are toxic or labile. Complexes can also (v) provide a means of co-regulating pathway enzymes or ensuring proper enzyme dosage and degradation.[403,404] An example of the co-regulatory properties of multimerization was found by Singh et al. with the discovery of a regulatory subunit in yeast FAS that is able to modulate functional properties of the enzyme's subunits.[128]

The notion that substrate channelling enhances the kinetics of the pathway has, however, a flaw, which is the fact that the diffusion of metabolites in the cytosol is faster than the catalytic rates of non-perfect enzymes. Thus, diffusion should not be a limiting factor in non-channeled metabolic pathways.[397,405] The only scenarios in which diffusion of the metabolites is rate-limiting are (i) in enzymes with high enough k_{cat}/K_M ratios for the diffusion to become the limiting step of the pathway[406] and (ii) in cases where the concentration of the enzyme is very low.[397] However, it has been demonstrated that, for some complex metabolic pathways, such as the Krebs tricarboxylic

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

acid cycle, it is essential to maintain high local concentrations of substrate and intermediates since average concentrations are insufficient to maintain the measured rates.[407] The very high concentrations of type II FAS enzymes, discussed below, are well in line with the observation that concentration is crucial for reaching the needed efficiency.

As for megasynthases, where type I FAS is included, their architecture not only facilitates substrate transfer between the different domains, due to the physical proximity of the catalytic domains, but also has a tethered carrier (ACP) responsible for taking the product of one domain to the next in the pathway. ACP-mediated substrate shuttling prevents the loss of the acyl moiety that is covalently bound to ACP and increases the catalytic efficiency by maintaining high local concentrations of the ACP-attached substrates and short diffusion distances between the catalytic centers.[53] Thus, high local concentrations of intermediates are combined with high local concentration of active centers, creating a powerful, synergistic concentration effect.

The megasynthase architecture also seems to allow the cell to produce fewer units of the protein for its metabolic needs, saving resources. For instance, in *E. coli*, due to the lower organizational level of the type II FAS system, the enzymes involved in fatty acid synthesis need to be present in high concentrations, becoming the third most abundant class of proteins.[408,409]

Within this context, the following section deals with several standing hypotheses that try to shed some more light into this matter.

4.2. Role of the ACP in Enzyme Efficiency

ACPs are ubiquitous proteins found in all kingdoms of life. They are involved in metabolic pathways, shuttling acyl molecules for, primarily, fatty acid and polyketide biosynthesis, and non-ribosomal peptide biosynthesis.[196,410,411] In the context of fatty acid synthesis, ACP is responsible for the efficient and timely shuttling of the chemical intermediates between the active sites of the domains/enzymes involved in fatty acid chain elongation.[412−414] The transport mechanism employed by ACP relies on the PNS cofactor to hold the different intermediates covalently bound to the ACP domain. This prosthetic arm of 18 Å establishes bonds in a covalent yet transient fashion, through a reactive, labile thioester bond.

ACP, as the FAS system itself, is classified into type I, as is the case of the animal or fungi ACP, attached to a megasynthase through a flexible tether,[25,415] or as type II, as a freestanding, independent monomer involved in fatty acid and polyketide synthesis in bacteria, plant and algae plastids, and mitochondria.[416] In both classes, ACP is responsible for shuttling the nascent fatty acids from active site to active site.

Within type I ACP, two alternative solutions to the substrate shuttling process exist. The fungal and CMN bacterial FAS is shaped as a rigid barrel, maintaining the ACP physically constrained inside.[57,126] In the animal type I FAS, the ACP domain is located next to the KR domain in the FAS polypeptide, after a 35 Å long flexible linker of 10 residues, and freely roams in the outer periphery of the enzyme.[25]

The best-studied type II bacterial ACP belongs to *E. coli*—it is highly acidic (p*I* of 4.1) and extremely soluble, and it interacts with more cellular proteins than any other monomeric protein known.[417] Its presence makes up to 0.25% of *E. coli* cytosolic protein content, suggesting high concentration.[418,419] Type II ACP not only transports the substrate through the cytosol, but also shields it from the solvent in its hydrophobic pocket, and delivers it to the correct catalytic partner.

As the flexible linker present in type I ACP physically constrains animal ACP to remain near the FAS scaffold, the substrate remains near the active sites where it is supposed to bind. This cancels out the need to sequester the substrate altogether. On the other hand, without dynamic conformation transitions induced by ligand binding, the mechanism through which ACP correctly and, eventually, orderly transports the substrate to the intended catalytic domain remains unknown.

Besides the strategy used by animal ACP to drive the intermediates into the adequate active sites in an orderly fashion, the ACP-mediated intermediate transport mechanism is itself far from clear. Most anabolic pathways leave the intermediates free to roam in the medium, and simple diffusion leads them from enzyme to enzyme, a simple scheme that works many times perfectly well and does not introduce a kinetic burden.

However, considering that the whole catalytic cycle of FAS includes a high number of catalytic reactions (7 condensations and 21 β-carbon processing steps),[25] the presence of ACP is a massive help comparatively to the alternative diffusion of the intermediates in the cell medium. Even if the diffusion itself is not rate-limiting in each individual step of the pathway, the large number of product releases/substrate entrances that the pathway comprehends might slow down the global speed of the pathway without ACP.

Another role ACP might have is to protect the intermediates from collateral reactions.

Since the transformation required to produce a fatty acid is so complicated, the likelihood of secondary reactions in the middle of the long process is high, which would justify the need to protect the reactive groups during the extensive process of palmitic acid formation. It is also possible that collateral reactions might lead to cytotoxic products, presently unknown, which evolution precluded the formation through the mechanism of intermediate protection provided by ACP. However, it has to be noted that, in numerous other anabolic processes, the products are biosynthesized by long enzymatic cascades and the intermediates diffuse through the solvent, from enzyme to enzyme, without the need for protection.

It might be also hypothesized that ACP's role is to improve the solubility of the multi-enzyme cycle intermediaries. Since the intermediates are very hydrophobic but always keeping an amphiphilic character, with an affinity for cell membranes, it might happen that, throughout the anabolic process, the intermediates migrate to the cell membrane. Implementing an anabolic route with intermediates captured and returned to the membrane seems a rather complicated scenario. In this sense, the ACP would serve to prevent the intermediaries from migrating to the membrane.

Finally, one might hypothesize that the ACP purpose is to improve the multi-enzyme kinetics through a concentration effect. The covalent tethering of ACP to the type I FAS scaffolds increases the local concentration of intermediates, favoring the reaction through a simple intermediate-concentration effect. This hypothesis is consistent with the emergence of a megasynthase architecture, which also increases active sites' concentrations. In fact, considering the amplitude of the reactional space, and the number of catalytic domains enclosed in the same space, the fungal FAS accounts for a concentration of 1.8 mM of active sites. Similarly, the animal FAS has a concentration of 1.2 mM of active sites.[408] This difference might be responsible for the turnover rates observed in type I systems:

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

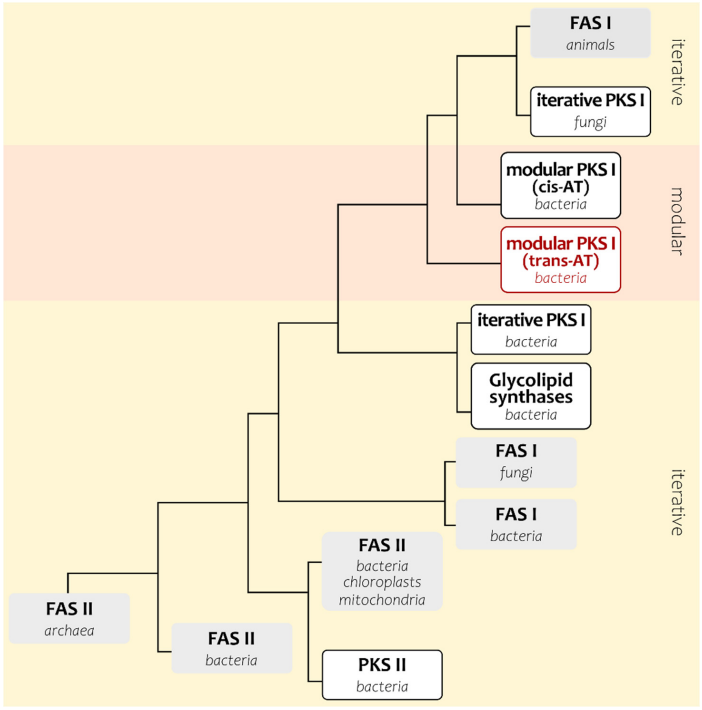


Figure 19. Simplified representation of the phylogenetic/evolutionary relationship between PKSs (in white) and FASs (in gray). Most relationships were derived from Bayesian statistics of acetyltranferase domain (AT) sequences and ketosynthase (KS) domain sequences representing all classes of PKSs and FASs as determined by Kodama and co-workers.[110] The particular case of modular type I trans-AT PKSs is highlighted in red, as alternative models propose that these PKSs can result from acetyltransferase (AT) loss in modular type I cis-AT PKSs;[423] hence, type I trans-AT PKSs interact with freestanding AT enzymes to load acyl primers/extenders.[280,428]

the fungal system is highly efficient completing 18.2 cycles of palmitic acid production per second,[420] whereas animal (chicken,[421] human[295]) and CMN bacteria (*Corynebacterium ammoniagenes*)[422] complete 3.4, 2.7, and 2.0 cycles per second, respectively. Together, a high concentration of intermediates and active sites might significantly affect enzyme efficiency.

In the case of the type II FAS—where catalytic domains are organized as individual enzymes freely diffusing in the cytosol—the compensating mechanism for the lack of an efficient intermediate-delivery system seems to be an increase in the copy numbers of enzymes, increasing the total concentration of active sites. A clear example is that of *E. coli*'s FabD (malonyl transferase, 10 000 copies) and FabA (dehydratase, 23 000 copies), which are the third most abundant proteins in *E. coli*.[409] The "concentration effect" is an attractive explanation for the emergence of an ACP tethered to a multi-enzyme during evolution. Despite this, substrate transporters tethered to multi-enzymes represent unusual protein architectures. Nature has privileged other alternative strategies to increase enzymes' catalytic power in most other anabolic circuits, such as lowering the limiting chemical barrier or lowering the product's affinity.

5. RELATION OF FAS WITH PKSs AND THE SYNTHESIS OF SECONDARY METABOLITES

PKSs are a family of enzyme systems that share an evolutionary history as well as large structural and functional similarities with FAS systems. Many lessons learned studying FAS systems could be translated to the PKSs and vice versa, which motivates the parallel study of both. Therefore, it makes sense to shortly summarize the state-of-the-art knowledge on the PKS systems and point out how its understanding promotes the progress in the understanding of the FAS systems and their engineering to synthesize highly complex chemicals of our desire. The interested reader can find more in-depth information about the structure, function, and biotechnological application of PKSs in several excellent reviews available in the literature.[111,234,423−428]

5.1. PKSs

The PKSs are involved in the secondary metabolism that, similarly to the FAS family, catalyzes condensation of acetyl- and malonyl-CoA primary metabolites to synthesize β-ketoacyl polymers. Like FASs, the PKS systems are also organized as sets of monofunctional enzymes (type II PKS) or multi-domain megasynthases (type I PKS), the latter still divided into FAS-like iterative multi-enzymes and non-iterative modular multi-

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

enzymes, which resemble an assembly of iterative PKSs that lost the ability to run iteratively, instead translocating the intermediate to the downstream module.[110,234,424,427] Unlike FASs, the PKSs catalyze the synthesis of a large diversity of polyketides, with varying carbon reduction states and degrees of aromaticity.[429−433] This polyketide diversity is achieved because, while the condensation step catalyzed by AT and KS is conserved in both FASs and PKSs, the KR, DH, and ER carbon processing domains/enzymes are selectively present or absent in PKSs,[110,434] such that the ketide extender groups can be converted into alcohols, alkenes, or alkanes.[234,424,427] Additionally, some PKSs can also accept acyl-CoA starters, other than acetyl- and malonyl-CoA characteristic of the FAS systems,[435,436] use other elongation substrates (e.g., methylmalonyl- and ethylmalonyl-CoA), or include catalytic domains/enzymes that do not exist in the FAS systems, namely methyltransferases or aromatases, which broadens the variety of synthesized polyketide secondary metabolites.

Alongside the discovery of an increasing number of PKS-encoding genes, several of their resulting products were also identified to exhibit potent biological effects, namely as antibiotics, antitumor compounds, natural insecticides, or immunosuppressive agents.[433] Consequently, the last decades witnessed increasing attention focused on the characterization of a growing number of PKS-synthesized natural products and, more importantly, the exploration of the potential of the PKSs for the *de novo* synthesis of novel compounds with applicability in health and technology.

5.2. Evolutionary Relationship between FASs and PKSs

The highly conserved nature of the KS domains, the phylogenetic analysis of polyketide megasynthases from bacteria and fungi, and the comparison between distinct types of PKSs and FASs point out that animal FAS are probably phylogenetically related to the iterative type I PKSs from bacteria (Figure 19).[110] X-ray structure studies by Maier et al., where the domain organization of animal FAS was compared with that of modules 1 and 4 of the type I modular PKS 6-deoxyerythronolide B synthase (DEBS), further confirmed that, at both the sequence identity and domain organization levels, animal FAS resembles more closely type I modular PKSs than it resembles any other FAS system.[108]

Recently, the evolutionary origin of the singular domain organization of the 2.6 MDa fungal FAS was portrayed by Bukhari et al. on the basis of a combination of phylogenetic analysis and X-ray crystallography characterization of the KS and ER domains,[403] as these were observed to establish strong protein:protein interactions connecting the dome and central wheel regions of fungal FAS.[56] The study identified gene insertions in the genes codifying a type II KS FabY and a DfnA-ER enzyme of a type I modular PKS DfnA, which resembled the KS- and ER-encoding regions of the fungal FAS gene. Upon subsequent structural analysis, the corresponding insertions were found to promote protein:protein interactions in FabY and DfnA-ER dimers and fungal KS and ER domains.[403,437] It was thus concluded that the KS and ER domains of fungal FAS derived from the type II bacterial FabY and the type I DfnA-ER PKS.[403] The evolutionary profile of the remaining domains of fungal FAS remains unclear, but the protein structure of DH seems to resemble the observed one in PKS-like modular megasynthases in polyunsaturated fatty acid biosynthesis.[403,438]

While it is clear that the "natural histories" of FASs and PKSs are closely related, phylogenetic analysis and transcriptome surveys of PKSs in both *protista* and *eukarya* have been suggesting that, while the FAS genes seem to have evolved to conserve enzyme function and the specificity toward specific long-chain fatty acids, the PKS gene families are highly expanded across different families of organisms, as a result of a large number of gene duplications and gene losses,[110,439−441] leading to a greatly diverse synthesis of polyketides.

5.3. Polyketide Synthesis and the PKS Machinery (Megasynthases vs Monofunctional Synthases)

As in FASs, PKS systems are also constituted by megasynthases (type I systems) and monofunctional enzymes (type II systems).[433] A particular class of PKS systems (type III) has also been identified, which encompasses a family of homodimeric KSs that does not make use of the ACP-transporter to perform condensation reactions of polyketide building blocks, directly using acyl-CoA building blocks instead.[442]

Type I PKS systems encompass the megasynthase multienzymes. These share architectural features with animal FAS, namely its homodimeric assembling. This broader class of PKS systems is the most complex among all PKS system types, and it is present in every fungus and most bacteria.[110] It can function iteratively, with multiple elongation cycles within a single multienzyme cycle, similarly to the iterative animal FAS, or in large multi-modular PKS assemblies, in which polyketide synthesis occurs downstream a series of non-iterative PKS modules interacting covalently or non-covalently. In the latter modular organization, each module has a condensation part, composed by an AT and a KS domain, and a variable processing part, responsible for the differentiation of the C_β of the elongating unit.[423,443] Consequently, type I PKS systems can synthesize saturated, unsaturated, and aromatic, linear and cyclic polyketides, depending on domain assembling.[234] Moreover, they can also form "dynamic hybrid complexes" either with individual enzymes of the type II PKS system or with non-ribosomal peptide megasynthases.

Type II PKS systems, only found in bacteria so far,[110] synthesize a large family of highly functionalized polycyclic aromatic compounds, among them some first-line antibiotics, anticancer drugs, veterinary medicines, and agrochemicals,[431,444,445] rendering these enzymatic systems quite appealing. Most of what is known about this PKS class is a result from the large body of work conducted in the actinorhodin biosynthesis pathway of *S. coelicolor*.[424] The initiation and condensations steps of type II PKSs are very much like those of type II FASs: the ACP protein carries the starter and the extender units from the MAT enzyme to the KS enzyme (homologue of FabD-H or FabB/FabF enzymes of type II FASs), where polyketide elongation occurs. However, the carbon processing stage differs significantly from FASs. Aside from the above-mentioned higher promiscuity of the KS enzyme of the type II PKS system toward starter units for polyketide synthesis, three main differences arise: (i) type II PKS KS enzyme is heterodimeric (opposite to the homodimeric FabH); (ii) only the KR and, possibly, DH enzymes (homologues to FabG and FabA/FabZ enzymes in type II FASs) were identified in the type II PKS system; and (iii) the synthesis of the final aromatic polyketide products relies on the activity of an aromatase (AR) domain,[446] which finds no equivalent in the FAS systems, and a final TE enzyme that functions as both a TE and a cyclase (CYC-TE), and it is responsible for final product cyclization and release.[446,447]

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

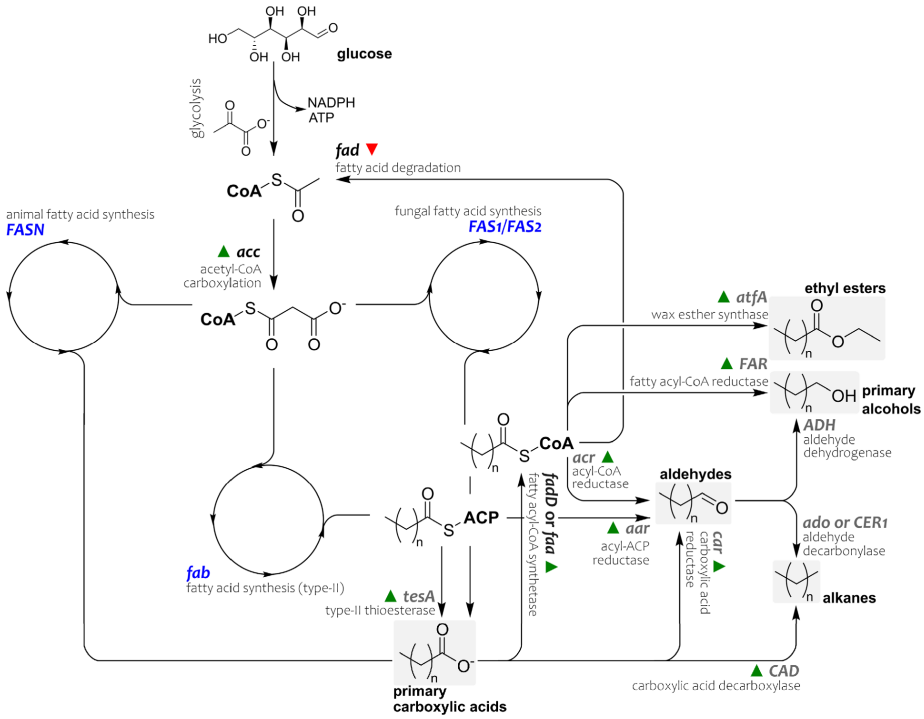


Figure 20. Common metabolic engineered pathways for production of fatty acids, biofuels, and commodity chemicals (primary carboxylic acids,[116,468,479,480] alkanes,[389,481,482] and primary alcohols,[468,480,481,483] aldehydes,[480] and ethyl esters[468,484]), using glucose. Upon the glycolysis of glucose to produce energy in the form of NADPH and ATP, the pyruvate product is decarboxylated by pyruvate decarboxylase to form the acetyl-CoA primer, which can be used as a primer of fatty acid synthesis or a substrate for the synthesis of malonyl-CoA extenders. Genes encoding enzymes of fatty acid synthesis are written in blue for the animal (*FASN*), fungal (*FAS1/FAS2*), and bacterial (*fab*) pathways, while genes encoding enzymes leading to desired end-products are written in gray. Genes commonly upregulated or downregulated by metabolic engineering are represented with ▲ or ▼ arrows.

In addition to type I and type II PKSs, most recent phylogenetic studies indicate that all land plants, and a few lineages of bacteria and fungi, synthesize a distinct type III PKS family of enzymes.[448,449] Unlike other PKSs or FASs, type III PKS systems do not rely on ACP to perform acyl-CoA supply to KS enzymes. Instead, they are KS homodimers that bind different CoA starters (e.g., benzoyl-CoA[450] or *p*-coumaroyl-CoA[451,452]) and catalyze several successive condensation reactions of malonyl- or acetyl-based building blocks,[453,454] based on the size of their active-site cavity. The resulting products are highly oxidant compounds that are prone to cyclization by either Claisen or aldol condensation and lactonization, after tetraketide formation,[455] and often constitute important precursors in the synthesis of functional compounds involved in protection, resistance, or differentiation mechanisms.[448,449]

5.4. Short Reference to the Engineering of PKSs

As the knowledge on the structure and function of PKSs expanded, the idea of making use of PKSs for *de novo* synthesis of polyketides became increasingly popular. While few works succeeded in catalytic reprogramming fungal type I iterative PKSs,[456,457] bacterial modular PKSs have been the most regarded for enzyme engineering in the past few decades.[427,429,458]

Early rational production of new polyketides was carried out by *in vivo* selective expression of type II PKSs from a set of biosynthetic pathways in *Streptomyces* strains.[459] In this regard, the characterization of the function and specificity of several type II PKSs was key to outline a set of strategies for the rational design of polyketides using characterized type II PKSs.

Following similar reasoning, and even though attempts to explore bacterial modular PKSs through domain swapping are less frequent in literature,[427,429] most common strategies focus on the engineering of synthetic PKS genes through module swapping, in order to produce varying chimeric PKSs and seek characterization of resulting products.[460,461]

While module swapping is a common engineering strategy for modular PKSs, a more common strategy for enzyme engineering is the use of directed engineering at specific amino acid positions. The approach has been already employed in the engineering of type III PKSs for the production of aromatic polyketide products, as type III PKSs function as homodimers that do not require additional enzymes for polyketide synthesis.[462] Alternatively, and particularly for multi-enzymes,

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

inactivation of specific domains by mutagenesis has also been suggested for control of the polyketide reductive cycle, as an alternative to the more cumbersome domain deletion.[460,461]

With the structural characterization of module assembling in modular PKSs, further *in vitro* characterization of PKSs and their reaction products, and the rapid advancements in bioinformatics tools and phylogenetic analysis algorithms, great avenues can be expected in the field of PKS engineering and new polyketide synthesis.

In summary, the PKS systems share a high similarity and a close evolutionary relationship with the FAS systems. They have been regarded as natural synthetic machines, which can be engineered to produce high value chemicals, whose structural complexity and stereochemistry render their synthesis in solution extremely challenging, if not impossible, with current means. Many lessons can be taken for the engineering of FASs from what has been achieved in the PKS world, in particular from the engineering based on specific amino acid mutations. A synergy between research of both synthetic systems is a key for the development of FAS systems that synthesize the compounds of our choosing. The next section will delve in the progresses on this field.

6. BIOTECHNOLOGICAL APPLICATIONS OF FASs

Recent years have witnessed an increasing interest in the engineering of FASs and its related metabolic pathways for biofuel production by microorganisms, because fatty-acid-based fuels exhibit high energy density, low hygroscopicity, miscibility with diesel fuels, reduced purification costs, and compatibility with existing infrastructure.[463] However, the resulting products of the engineered variants are commonly a mixture of fatty acids, with varying physicochemical properties depending on their elongation, branching pattern, and unsaturation level.[206,464] Nevertheless, with the development of novel and more efficient methods to assess the composition of fatty acid mixtures,[465,466] and the development of more specific organisms/biosystems for fatty acid production, the metabolic engineering of fatty acid synthesis as well as the engineering of individual FAS enzymes/domains can soon become a sustainable strategy for the production of biofuels and commodity chemicals, and for novel drug synthesis.[467,468]

6.1. Metabolic Engineering of FASs

Metabolic engineering has been probably the most explored alternative to produce short- and medium-chain fatty acids and oleochemicals.[469−473] To date, *E. coli* remains one of the most popular hosts for metabolic engineering,[474−476] even though success has been hampered by the fact that the products of bacterial FAS are fatty acyl-ACPs instead of fatty acyl-CoAs,[471] and that bacterial cultures are often susceptible to phage attacks.[477] On the other hand, yeast systems, namely the industrial workhorse *S. cerevisiae* or *Yarrowia lipolytica*, are also popular alternatives and have been specifically pointed out as potential platforms for synthesis of drop-in transportation fuels and oleochemicals.[392,473,478] However, the current ability to engineer the eukaryotic type I metabolism is not as advanced as is for the prokaryotic type II metabolism.[468]

6.1.1. Metabolic Engineering in Bacteria. The overall primary sources in fatty acid biosynthesis are acetyl- and malonyl-CoA, which in turn can be obtained from the degradation of simple sugars by the glycolysis pathway. Hence, common metabolic engineering strategies target genes responsible for the acetyl-CoA carboxylation (*acc*) to increase the supply of malonyl-CoA,[468,485−487] and attempt the down-regulation of the fatty acid degradation (*fad*) genes to prevent fatty acyl-CoA *in vivo* consumption (*acc* and *fad* genes in Figure 20).[116,468,479] In particular, the deletion of *E. coli fad* genes, leading to fatty acid oxidation and enhanced fatty acid synthesis, was combined with the expression of a mutant TE domain with higher specificity for smaller fatty acid-ACPs,[488] to produce non-natural 9 to 14 carbons long fatty acids[389,476] A large diversity of fatty acids, from 4 to 18 carbons long, was also produced in *E. coli* host cells, upon insertion and expression of a standalone animal TE-II enzyme and an animal FAS (see *tesA* and *FASN* genes in Figure 20).[489] These short- and medium-chain products are particularly important products in the production of bio-based fuels that may replace petrol-based ones.[490]

Further expression of fatty acid biosynthesis (*fab*) genes, namely the *fabD* (encoding the FabD enzymes, homologous to the MAT domain) or the combined *fabG* and *fabZ* (encoding the FabG and FabZ enzymes, homologous to the KR and DH domains), was observed to result in an increased production of long-chain fatty acids, namely palmitic and stearic acids,[491,492] while overexpression of the *fabH* gene (encoding the FabH enzyme, homologous to the KS domain), led to a higher production of non-natural short- and medium-chain fatty acids upon expression of a modified single mutant TE enzyme.[389] In particular, the metabolic engineering combining the over-expression of *fabH* with the sequential operation of engineered fatty acyl-CoA synthetase, fatty acyl-CoA reductase, and fatty aldehyde decarbonylase resulted in the production of 9- to 14-carbon alkanes with yields of up to 580.8 $mg{\cdot}L^{-1}$.[389] The overexpression of another *fabH* gene encoding a FabH enzyme with specificity toward branched-chain acyl-CoA starters was also observed, in combination with the upstream α-keto acid biosynthetic pathway, to produce high titers of elongated branched-chain acyl-CoA.[493] The synthesis of the latter compounds is particularly relevant to improve the cold-flow properties of biofuels, as linear hydrocarbon-based biofuels operate only during warm weather.[494]

An international patent was also published in 2017, where *E. coli* host cells were further engineered to improve the production of branched-chain acyl-CoAs in detriment of linear short-chain ones, and finally achieve their conversion into the respective alcohols and esters.[495]

Despite the fact that KS and TE enzymes were pointed out to be particularly important at defining the narrowness of bacterial fatty acid product profiles, joint mechanistic kinetic models and compositional studies of *E. coli* type II FAS indicated that varying ratios in domain expression can also alter product profiles by substrate pool availability.[496]

In an alternative approach, based on the modular organization of several PKS genes in bacteria, a modular approach—where genes from the carbon metabolism of *E. coli* were assembled into an acetyl-CoA formation module, an acetyl-CoA activation module, and a malonyl-CoA consumption module—was shown to yield a 20-fold improvement in fatty acid synthesis in shake flask cultures and an 8.6 $g{\cdot}dm^{-3}$ fatty acid synthesis rate on a 20 dm^3 fed-batch fermenter, given that optimized expression and culture conditions were met.[475] The study emphasized that the combinatorial optimization of the transcriptional levels of the devised modules could overcome metabolic pathway bottlenecks that often interfere with long-term cell viability and, consequently, improve fatty acid production yields. On a side note, moderate improvements in fatty acid production were also observed by gene fusion of the two FAS genes encoding the α

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

and β subunits of fungal FAS,[497] which again supports that selective gene fusion can improve fatty acid yields, in combination with both rational and metabolic engineering.

6.1.2. Metabolic Engineering in Yeast. Metabolic engineering of *S. cerevisiae* strains by overexpression of primary genes of fatty acid synthesis (*FAS1/FAS2*) and enzymes, such as TEs (*tesA*), wax ester synthases (*atfA*), or fatty acyl-CoA reductases (*FAR*), led to the production of fatty acids, fatty acid ethyl esters, or fatty acid alcohols from simple sugars (Figure 20).[468,498] The resulting ester and alcohol products, in particular short-chain ones, can find broad applicability in industry, namely for the production of flavors, fragrances, solvents, paints, or coatings.[499,500]

In a different approach, the combination of the swapping of the malonyl-palmitoyl-transferase (MPT) domain from the FAS of the yeast *Y. lipolytica* by smaller TE domains, with metabolic engineering of its lipogenic phenotype, resulted in the production of medium-chain fatty acid products of 12 or 14 carbons with yields up to 30% higher than the native FAS pathway, together with an improvement in the production of fatty acids, fatty acid ethyl esters, or fatty acid alcohols.[392] Since the MPT domain is responsible for both palmitoyl release and malonyl uptake, the origin of medium-chain fatty acid production remains elusive. In another study, the action of a carboxylic acid reductase in an engineered *S. cerevisiae* FAS capable of synthesizing hexanoic and octanoic acid, resulted in the production of octan-1-ol with yields of 49.5 $mg{\cdot}L^{-1}$, even though product toxicity could not be circumvented to further increase the yield (please see the metabolic pathway described by genes *FAS1/FAS2*, *tesA*, *car*, and *ADH* in Figure 20).[480] In this regard, the deletion of a peroxisome assembly factor in *Y. lipolytica* led to cell growth and a ca. 50-fold increase in primary alcohol production, e.g., decan-1-ol, which suggested that fatty alcohol catabolism by the peroxisome could be responsible for the observed toxicity of fatty alcohols.[501] Such alcohol products are pointed out as promising replacements for diesel and jet fuels because they promote lower emissions of NO, NO_2, and CO pollutant gases and result in enhanced engine performances.[502,503]

Alternative to metabolic engineering approaches, a recent multi-dimensional approach combining adaptive laboratory evolution, site-directed mutagenesis, and metabolic engineering was able to further improve the yield of 8- and 10-carbon medium-chain fatty acid production by engineering the FAS genes of *S. cerevisiae* and *Mycobacterium vaccae*, resulting in an improvement of 250-fold over starting strains ($1.39\ g{\cdot}dm^{-3}$).[478] In order to overcome the toxicity of these products by accumulation in the organism,[504] directed evolution was also successfully employed to evolve a 10-carbon long fatty acid specific transporter that, when combined with other techniques, could ensure sub-inhibitory product concentrations, further stimulating the production of the desired fatty acids.[478]

Although less common, the engineering of the fatty acid biosynthetic pathway together with the expression of specific enzymes can also be combined to perform asymmetric organic synthesis, which represents a more appealing solution to most common biosynthetic pathways that make use of lipases, where resulting racemic mixtures are often composed of equal amounts of their racemates.[505,506] In that regard, the selective overexpression of FAS, an aldo-keto reductase, and an acetoxy ketone reductase in *S. cerevisiae* showed improved stereoselectivity in the reduction of β-keto esters.[507] Later, a recombinant *S. cerevisiae* overexpressing FAS and a glucose dehydrogenase was observed to produce enantiomeric excesses above 80% in chiral alcohols, which are common constituents of pharmaceuticals, flavors, and fragrances.[508] However, the development of a general approach based on the use of whole yeast showed to be more difficult to achieve, as the identification of a complete set of reductases catalyzing alternative other enantioselective reductions would require exhaustive knockout and genomic studies.[507]

6.1.3. Orthogonal Fatty Acid Biosynthesis. Orthogonal fatty acid biosynthesis was achieved by insertion of a bacterial type I FAS in *E. coli*, in order to overcome the tight regulation of the same pathway by the native type II FAS (*FAS1/FAS2* and *fab* pathways in Figure 20).[483] Despite improving the fatty acyl-CoA production, the final yields in fatty acids were presumably still hampered by the rate-limiting action of fatty acyl-ACP from type II FAS on the malonyl-CoA supply, which relied on acetyl-CoA carboxylation by the acetyl-CoA carboxylase encoded in *acc* genes.[483,509] To overcome this limitation, the combination of the former strategy with an optimal expression of enzymes in the alkane biosynthesis pathway, specifically an acyl-ACP reductase catalyzing acyl-ACP reduction to acyl aldehyde, and an aldehyde decarbonylase catalyzing the decarbonylation of the resulting acyl aldehyde to alkane, bypassed the feedback inhibition mechanism resulting from acyl-ACP accumulation, and led to promising final titers of 100 $mg{\cdot}dm^{-3}$ of long-chain alcohol and alkane production (*aar* > *ADH* and *aar* > *ado* in Figure 20).[481] The heterologous expression of a similar combination of enzymes in *E. coli* had already shown increases in the production of 13- and 17-carbon mixtures of alkanes and alkenes in *E. coli*, with titers over 300 $mg{\cdot}dm^{-3}$.[482] In a similar approach in *S. cerevisiae*, where a heterologous type I bacterial FAS (analogous to *FAS1/FAS2* in Figure 20) was co-expressed together with a fatty acid acyl-CoA ligase and a wax ester synthase/acyl-coenzyme A:diacylglycerol AT (*faa* and *atfA* in Figure 20), a 6-fold improvement was observed in the production of 16- to 18-carbon fatty acid ethyl esters.[484] Further improvements would require manipulation of acetyl-CoA levels and further elimination of competing fatty acid-utilizing pathways.[510−513] At least two patents were published in recent years, in which heterologous gene expression and orthogonal fatty acid biosynthesis have been combined for improved synthesis of specific fatty acid profiles and further conversion into aldehydes, alcohols, or esters.[514,515]

6.2. Enzyme Engineering of FASs

6.2.1. FAS Engineering for Biofuel Production. The yeast FAS of *S. cerevisiae* has been the preferred target for enzyme engineering, as it is the most kinetically competent out of all FASs.[516,517] Despite its efficiency, FAS megasynthases are quite specific for the formation of longer fatty acids, namely palmitic or stearic acids (Figure 21).[279,485] On the other hand, while slower, expression of bacterial FASs has been shown to produce fatty acids of varying chain elongations.[518,519] Hence, merging the kinetic efficiency of FAS megasynthases with the genetic diversity of *fab* bacterial genes, is a very appealing alternative for FAS megasynthase engineering. The heterologous integration of bacterial enzymes, particularly ester hydrolases, into the central wheel of yeast FAS has been a successful approach to alter fatty acid chain specificity (Figure 21).[392,520]

In fact, the expression in *S. cerevisiae* of a fungal FAS from *Rhodosporidium toruloides*, where a second ACP domain was replaced by a short-chain specific bacterial TE domain from

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

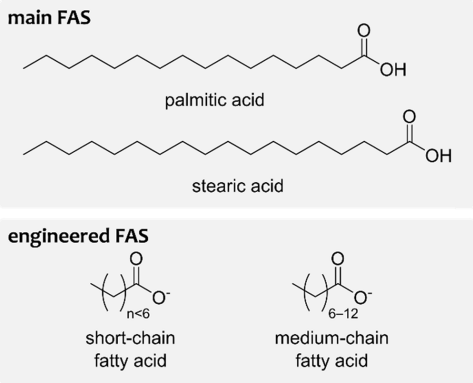


Figure 21. Representation of the main fatty acids synthesized by native FAS in fungi and animals, and the short- and medium-chain fatty acids (with most applicability in biofuel production) currently synthesized from the engineered FAS strains.

Acinetobacter baylyi,[521] resulted in a 20-fold increase in the production of hexanoic and octanoic acids relatively to the wild-type form.[520] The switching of the MPT domain by a small TE domain in the fungal FAS of *Y. lipolytica* has also shown to alter substrate specificity and increase medium-chain fatty acid production in up to 9.67 $g \cdot dm^{-3}$ in a 3 dm^3 bioreactor.[392] If additional rational engineering can be performed on the top of the commonly employed overexpression protocols, more potent enzyme catalysts may be designed. Such feat was already achieved for a TE domain from *E. coli*, in which a single mutant with 2-fold larger activity was engineered through a combination of error-prone polymerization condensation reaction and high-throughput screening assays.[522]

Rational mutagenesis at the active site of the KS and MPT/AT condensation domains of FASs also resulted in a differentiated production of fatty acids of varying elongation. In particular, limiting of fatty acid elongation by a double-mutant KS domain led to a substantial increase in hexanoic and octanoic acids production. When combined with a mutation in the MPT domain and the starting AT domain, to lower malonyl-CoA input against fatty acid export and to improve fatty acid initiation, an over 100-fold production of short fatty acids was observed.[523] These promising results culminated in an international patent describing a method to produce 6- to 12-carbon fatty acids (medium-chain fatty acids in Figure 21) in bacterial or fungus host cells.[524] In a study that quickly followed, the computer-aided identification of a residue interacting with palmitic acid in the KS domain of *Y. lipolytica*, led to the observation that mutagenesis of specific interacting residues by aromatic substituents can be used to control fatty acid elongation, even though this requirement was insufficient to produce a very efficient enzyme.[525] Only through metabolic engineering on the resulting mutant FAS, did the authors observe a significant increase in lipidic content.[465] Nevertheless, the study further confirmed that engineering of KS can team up with other engineering attempts to design efficient FAS enzymes with desired fatty acid product profiles.

The engineering of the MPT domain of yeast FAS could also be of particular interest, as the MPT domain is responsible for the high selectivity of this megasynthase.[47,57] However, since the control of fatty acid release seems to be particularly important for the yield of shorter chain fatty acid production, the modification of the bifunctional MPT domain in yeast FAS, which loads malonyl-CoA and unloads palmitoyl-CoA, may also compromise enzyme activity. In that regard, the engineering in yeast of a human FAS where the native TE domain was replaced by a short-chain specific TE domain, showed a 64-fold improvement in octanoic acid production relatively to the wild-type human FAS.[116] The study was also able to show that a covalently linked TE significantly increased the rate of the megasynthase in up to 9-fold.[116] Altogether, these results indicate that engineering of the animal FAS can not only be done in yeast, but it is also a more efficient approach for the production of short- and medium-chain fatty acids. Further optimization of the FAS-encoding gene led to additional improvement in fatty acid production; however only in combination with metabolic engineering aiming at disruption of fatty acid oxidation, a 181-fold improvement over the parent strain was observed.[526]

6.2.2. FAS Engineering for Polyketide Synthesis. While we have seen numerous examples in which FAS megasynthases can produce important fatty acyl-CoAs for biofuel production, they can also be manipulated to delve into the world of PKS engineering and polyketide synthesis. Both animal and fungal FAS have been used for this end, even though the animal FAS megasynthases are better characterized, as well as their composing domains.[104,315]

The idea that FASs could be manipulated to produce acyl-CoA precursors that could be of use for polyketide synthesis by PKSs soon raised interest. An early study performed in animal (insect) cells showed that the 3-hydroxybutyryl-CoA precursor of polyhydroxyalkanoate synthesis (Figure 22), could be obtained as the product of the first condensation cycle of a rat mutated FAS, in which the catalytic His878 of the DH domain

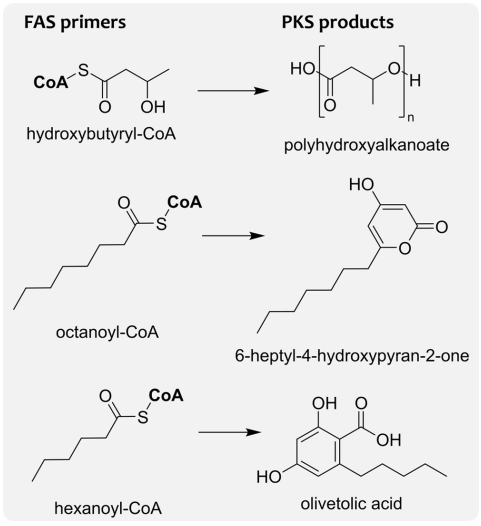


Figure 22. Some examples of primers that can be obtained from FAS engineering and the resulting polyketide products obtained by the action of PKSs or other engineered FASs.[527−529]

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

was replaced by Ala.[527] The desired polyhydroxyalkanoates products, which find use in the development of thermostable bioplastics, could then be produced by expression of a polyhydroxyalkanoate synthase.

A recent kinetic characterization of the transacylation reaction by the animal (murine) MAT domain toward several starter/loading acyl-CoA units, with varying length and volume, showed that it exhibits good efficiency for a broader range of substrates, even though the highest efficiency is observed for malonyl-CoA, in agreement with the expected function for FASs.[104] The fact that it also shows lower hydrolysis activity in comparison to known AT domains in PKSs,[212,530] also suggests it as a good candidate for PKS engineering. In fact, there are already reported examples in which PKSs can interact with MAT domains of FAS to initiate polyketide synthesis, namely in the biosynthesis of the romidepsin anticancer drug (FK228) by a type I modular PKSs,[531] or the biosynthesis of actinorhodin[532] or tetracenomycin[533] by type II PKSs (upper panel in Figure 23). More recently, the expression in microbial host cells of a fungal FAS, a tetraketide synthase, and an olivetolic acid cyclase, together with other supporting enzymes, resulted in the production of several cannabinoids and cannabinoid precursors and was published in an international patent (middle panel in Figure 23).[528]

Alternatively, taking advantage of the larger substrate specificity of several bacterial type II KS, it should also be possible to engineer enzymes and pathways comprising type II FASs to synthesize novel fatty acids, such as the less common branched fatty acyls, which could also be used as precursors for the synthesis of macrolactam antibiotics, ω-functionalized fatty acids, among other functionalized molecules, by other PKSs.[534] An example can be found in an international patent from 2016, where a type II FAS KS with specificity for an ω-functionalized acyl-CoA was expressed with a specific TE domain in *E. coli*, to produce α,ω-bifunctional fatty acids out of glucose (lower panel in Figure 23).[535]

In a bolder approach, genetic engineering of animal (murine) FAS in *E. coli* also indicated that it is possible to design PKS modules of varying reducing nature and inclusively design bimodular PKS assemblies, even though sequence optimization is required to prevent *in vivo* proteolysis.[111] Both these works emphasize that a better understanding of FAS structure and function represents important advancements in the development of biosynthetic-inspired industrial protocols.

Some examples of successful polyketide synthesis can be found from FAS engineering. In a recent study, a type I FAS from *C. ammoniagenes* was engineered into a reductive module capable of synthesizing the short-chain octanoyl-CoA, and a second non-reductive module producing the 6-heptyl-4-hydroxypyran-2-one polyketide from the former octanoyl-CoA, with an optimized yield of 35% (Figure 22).[529] The latter polyketide is a triacetic acid lactone that can be used as precursor for the synthesis of food preservative molecules, resins, adhesives, or plastics.[536] While for the first module, the authors again followed an approach of site-directed mutagenesis at the KS, MPT, and AT domains to improve the production of short-chain fatty acyl-CoAs,[523] in the second module, the activity of the KR domain was abolished by mutagenesis of the critical catalytic Tyr2227[537,538] by a Phe, and the priming of acetyl-CoA by the AT domain was impeded by mutating the catalytic Ser126 by Ala.[529] In another study in 2004, the same mutation on the KR domain of the yeast FAS of *Brevibacterium ammoniagenes* had already led to the production of the same 6-heptyl-4-hydroxypyran-2-one molecule, and this first emphasized the potential of FAS engineering for the benzene-free production of aromatic molecules.[539]

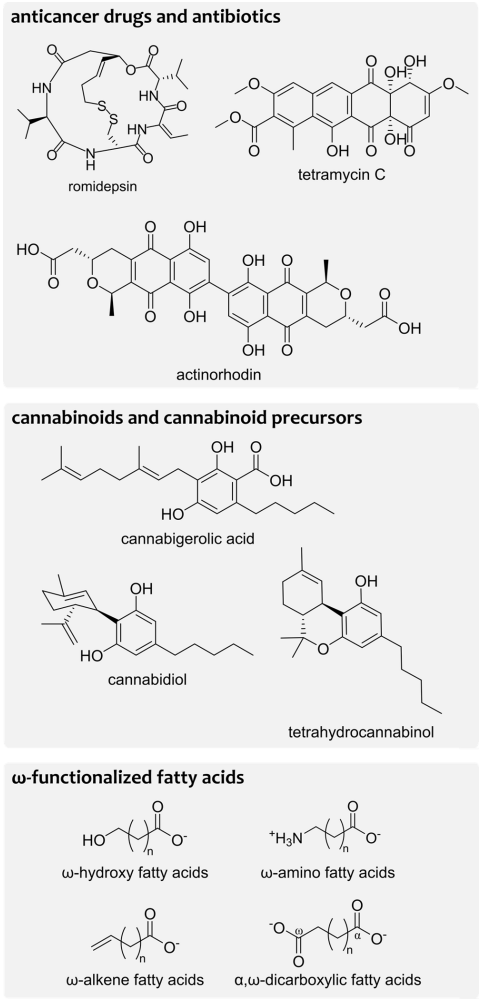


Figure 23. Some examples of products that could be obtained from the engineering of FASs, and upon combination with other PKSs: anticancer drugs, antibiotics, cannabinoids, and ω-functionalized fatty acids.

7. CONCLUSION

Animal FAS is a natural example of nanotechnology—it accommodates different substrates, carries very long and complicated chemistry, transfers its substrate from active site to active site through a shuttling arm, protected from the solvent, and synthesizes the desired product with remarkable accuracy, very much like a classical production line of a modern factory,

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

but at the nanoscale. We highlighted the current structural and biochemical knowledge on this enzyme. The progress has been tremendous, in particular in terms of structural elucidation. Whether the movement of ACP along the multi-enzyme cycle is random or driven, the determinants that control its path and sequential selectivity are of fundamental interest and constitute an area for future developments. This will imply moving from a static to a dynamic description of the structure of animal FAS.

We discussed the FAS chemistry critically, with minute detail. Most of the mechanisms are now understood, but gaps remain. The mechanisms of the KS and TE domains are less well understood than the others; a deeper understanding of the molecular basis of the TE product selectivity is also desirable. These are areas of intense research that promise to close a cycle of understanding in a not-far future.

We systematized the FAS's current biotechnological applications and altogether its potential. The synthesis of complex chiral molecules is a challenge for solution chemistry. The tackling of such a problem with biosynthesis, either for total synthesis or for specific synthetic steps, is a hot topic nowadays due to its synthetic capacity and because biosynthesis is green. FAS is and will be a valuable platform for tuned biosynthesis. As our understanding of the enzyme machinery expands, the capacity to use it to synthesize compounds of our choice will be facilitated.

Another promising field is the inhibition of FAS for therapeutic purposes. FAS is an exciting anti-cancer target. The side effects of FAS inhibition are not the typical cytotoxicity that most other drugs have. Overcoming its anti-nutritional side effects may deliver new drugs of extraordinary potential.

We formulated what we consider critical open questions throughout this Review and provided our opinions on the matter that are far from being definitive. Despite being impossible to cover the whole knowledge gathered on the animal FAS in the many decades of study, and being fair to the effort of so many researchers that worked on this subject, we tried to be comprehensive in organizing the current knowledge (and lack of it!) in a helpful way for both experienced researchers and beginners in the field, in the hope of catalyzing the quest for the many mysteries that this amazing megaenzyme still hides from us.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemrev.1c00147.

Video SI-1, depicting the biosynthetic pathway of the human FAS (MP4)

Table SI-1, a compilation animal FAS and its individual domains, the 3D structures of which are available from the Protein Data Bank; Figure SI-1, a representation of the dodecameric assembly of the barrel-shaped fungal FAS; and description of Video SI-1 (PDF)

AUTHOR INFORMATION

Corresponding Author

Pedro A. Fernandes − *LAQV, REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, 4169-007 Porto, Portugal;* orcid.org/0000-0003-2748-4722; Email: pafernan@fc.up.pt

Authors

Pedro Paiva − *LAQV, REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, 4169-007 Porto, Portugal;* orcid.org/0000-0002-7003-6055

Fabiola E. Medina − *Departamento de Ciencias Químicas, Facultad de Ciencias Exactas, Universidad Andres Bello, Autopista Concepción-Talcahuano 7100, Talcahuano, Chile;* orcid.org/0000-0002-0230-8717

Matilde Viegas − *LAQV, REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, 4169-007 Porto, Portugal;* orcid.org/0000-0002-4472-9691

Pedro Ferreira − *LAQV, REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, 4169-007 Porto, Portugal;* orcid.org/0000-0001-7015-5460

Rui P. P. Neves − *LAQV, REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, 4169-007 Porto, Portugal;* orcid.org/0000-0003-2032-9308

João P. M. Sousa − *LAQV, REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, 4169-007 Porto, Portugal;* orcid.org/0000-0002-4915-4946

Maria J. Ramos − *LAQV, REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, 4169-007 Porto, Portugal;* orcid.org/0000-0002-7554-8324

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemrev.1c00147

Author Contributions

⊥P.P., F.E.M., M.V., P.F., R.P.P.N. and J.P.M.S. contributed equally to this work.

Notes

The authors declare no competing financial interest.

Biographies

Pedro Paiva graduated in Biochemistry (2012) and holds a MSc degree in Biochemistry (2014), both from the Faculty of Sciences of the University of Porto. In 2017, he enrolled in the M2B - Medical Biochemistry and Biophysics Ph.D. programme (University of Lisbon, University of Porto and University of Coimbra) and is developing his PhD project at the High Performance Computing in Molecular Modelling group of LAQV/REQUIMTE. His main research area is Computational Biochemistry, with a particular interest in the study of catalytic mechanisms of enzymes with therapeutic potential, using Quantum Mechanics/Molecular Mechanics methods.

Fabiola E. Medina received her Master (2012) degree in Chemistry from the University of Concepción, Chile. In 2019, she completed her Ph. D. in Sustainable Chemistry, in the Theoretical and Computational Biochemistry Research group at the University of Porto, Portugal, under the supervision of Professor Pedro A. Fernandes. Currently, she is a researcher at the University of Andres Bello, Chile. Her scientific research focuses on applying different computational methods to understand biological processes, such as enzyme catalysis and protein dynamics. Her recent projects are focused on metalloenzymes, Metallo-Beta-Lactamase (MBL) and A Disintegrin And Metalloproteinase (ADAM) enzyme, both with implications in biomedicine.

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

Matilde Viegas holds a MSc degree in Biochemistry (2014) and a BSc in Biochemistry (2011), both obtained from University of Porto. In late 2017, she initiated her Ph.D. in Chemistry (Faculty of Engineering, and Faculty of Sciences at the University of Porto) dedicated to the study of the catalytic mechanisms of extremophilic enzymes with biotechnological applications in the biofuels industry. Since 2015, she has been working at the High Performance Computing in Molecular Modelling group of LAQV/REQUIMTE contributing to projects ranging from snake venom toxins to treatment of diabetes type 2. In 2017, she published the first homology model for the entire human FAS, including mobile domains previously absent in the crystal structure of mammalian FAS.

Pedro Ferreira is a PhD student in Chemistry in the University of Porto. His PhD project is being developed at the High Performance Computing in Molecular Modelling group of LAQV/REQUIMTE in the computational biochemistry field. The main focus of the project is the description of the catalytic mechanisms of enzymes, particularly coming from extremophile organisms, with potential for industrial applications using computational methodologies. In addition, the improvement of the catalytic rates of the studied enzymes through rational computational mutagenesis is also a goal. He has a MsC degree in Biochemistry, from the University of Porto obtained in 2014.

Rui P. P. Neves received his PhD in Chemistry in 2016, from University of Porto. He is now a hired PhD researcher at the High Performance Computing in Molecular Modelling group of LAQV/REQUIMTE. In his early research, he studied strategies to systematically parametrize manganese metalloproteins for subsequent extensive molecular dynamics, but started to study reaction mechanisms of oxidoreductases soon after. His current research interests focus on the study of enzymes with implications in sustainable processes for environmental remediation, chemical synthesis and energy production. In particular, he is exploring computational mutagenesis tools to rationally design enzyme mutants with improved activity for subsequent in vitro testing and pilot plant implementation.

João Sousa is graduated in Biochemistry (2016) and also holds an MSc degree in Biochemistry (2018) for the Faculty of Sciences of the University of Porto. In 2019, he received a PhD grant from the Fundação para a Ciência e Tecnologia, Portugal. His main research interest is the study of enzymatic catalysis using high accuracy computational methods. Since 2017, he has been studying the catalytic mechanism of enzymes by using mainly hybrid Quantum Mechanics and Molecular Mechanics methods.

Maria João Ramos is currently a Full Professor at the University of Porto, Portugal, interested in Computational Enzymology and Drug Discovery. She did a PhD in Quantum Chemistry at the University of Glasgow, and a post-doc in Molecular Modelling at the University of Oxford, UK. She was a Vice-Rector at the University of Porto, an Associate Director at the NFCR Centre for Computational Drug Discovery at Oxford and a Director at the Science and Technology Park of the University of Porto. She won several prizes throughout her career and, in 2014, was awarded an Honorary Doctorate by Stockholm University, Sweden.

Pedro Fernandes has a Ph.D. in Chemistry from the University of Porto, Portugal (2000) in Molecular Dynamics. Now a Professor in the Faculty of Sciences of the same university, he divides his time between teaching and research, mainly focusing on the elucidation of enzyme reaction mechanisms and enzyme engineering with computational QM/MM methods. He also works in the fields of computational protein structure, modeling, energetics, and mutagenesis. The chemistry of snake venoms is one of his specific interests. Drug discovery is also one of his essential fields. In his free time, he likes reading, travelling, diving, and doing sports.

ACKNOWLEDGMENTS

This work was supported by UIBD/50006/2020, PTDC/QUI-QFI/28714/2017, and PTDC/QUI-OUT/1401/2020 with funding from FCT/MCTES through national funds. P.P., M.V., P.F., and J.P.M.S. thank FCT/MCTES (Fundação para a Ciência e Tecnologia/Ministério da Ciência, Tecnologia e Ensino Superior) and ESF (European Social Fund) through POCH (Programa Operacional Capital Humano) and PORN (Programa Operacional Regional do Norte) for their Ph.D. grants, ref PD/BD/135268/2017, SFRH/BD/119206/2016, SFRH/BD/144205/2019, and SFRH/BD/131968/2017, respectively. F.E.M. thanks ANID FONDECYT 3200304. R.P.P.N. acknowledges funding from the project PTDC/QUI-QFI/28714/2017 through FCT/MCTES.

REFERENCES

(1) Fielding, B. A.; Frayn, K. N. Lipid Metabolism. *Curr. Opin. Lipidol.* **2003**, *14*, 389−391.
(2) van Meer, G.; Voelker, D. R.; Feigenson, G. W. Membrane Lipids: Where They Are and How They Behave. *Nat. Rev. Mol. Cell Biol.* **2008**, *9*, 112−124.
(3) Calder, P. C. Long Chain Fatty Acids and Gene Expression in Inflammation and Immunity. *Curr. Opin. Clin. Nutr. Metab. Care* **2013**, *16*, 425−433.
(4) Kendall, A. C.; Pilkington, S. M.; Massey, K. A.; Sassano, G.; Rhodes, L. E.; Nicolaou, A. Distribution of Bioactive Lipid Mediators in Human Skin. *J. Invest. Dermatol.* **2015**, *135*, 1510−1520.
(5) Bloch, K.; Rittenberg, D. Sources of Acetic Acid in the Animal Body. *J. Biol. Chem.* **1944**, *155*, 243−254.
(6) Klein, H. P.; Lipmann, F. The Relationship of Coenzyme A to Lipide Synthesis. 2. Experiments with Rat Liver. *J. Biol. Chem.* **1953**, *203*, 101−108.
(7) Klein, H. P.; Lipmann, F. The Relationship of Coenzyme A to Lipide Synthesis. 1. Experiments with Yeast. *J. Biol. Chem.* **1953**, *203*, 95−99.
(8) Wakil, S. J. A Malonic Acid Derivative as an Intermediate in Fatty Acid Synthesis. *J. Am. Chem. Soc.* **1958**, *80*, 6465−6465.
(9) Zhang, L.; Joshi, A. K.; Smith, S. Cloning, Expression, Characterization, and Interaction of Two Components of a Human Mitochondrial Fatty Acid Synthase. *J. Biol. Chem.* **2003**, *278*, 40067−40074.
(10) Brady, R. O. The Enzymatic Synthesis of Fatty Acids by Aldol Condensation. *Proc. Natl. Acad. Sci. U. S. A.* **1958**, *44*, 993−998.
(11) Gibson, D. M.; Titchener, E. B.; Wakil, S. J. Requirement for Bicarbonate in Fatty Acid Synthesis1. *J. Am. Chem. Soc.* **1958**, *80*, 2908−2908.
(12) Gibson, D. M.; Titchener, E. B.; Wakil, S. J. Studies on the Mechanism of Fatty Acid Synthesis. V. Bicarbonate Requirement for the Synthesis of Long-Chain Fatty Acids. *Biochim. Biophys. Acta* **1958**, *30*, 376−383.
(13) Wakil, S. J.; Titchener, E. B.; Gibson, D. M. Evidence for the Participation of Biotin in the Enzymic Synthesis of Fatty Acids. *Biochim. Biophys. Acta* **1958**, *29*, 225−226.
(14) Wakil, S. J.; Ganguly, J. On the Mechanism of Fatty Acid Synthesis. *J. Am. Chem. Soc.* **1959**, *81*, 2597−2598.
(15) Wakil, S. J.; Pugh, E.; Sauer, F. The Mechanism of Fatty Acid Synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **1964**, *52*, 106.
(16) Majerus, P. W.; Alberts, A. W.; Vagelos, P. R. The Acyl Carrier Protein of Fatty Acid Synthesis: Purification, Physical Properties, and Substrate Binding Site. *Proc. Natl. Acad. Sci. U. S. A.* **1964**, *51*, 1231−1238.
(17) Majerus, P. W.; Alberts, A. W.; Vagelos, P. R. Acyl Carrier Protein. Iv. The Identification of 4′-Phosphopantetheine as the

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

Prosthetic Group of the Acyl Carrier Protein. *Proc. Natl. Acad. Sci. U. S. A.* **1965**, *53*, 410−417.
(18) Alberts, A. W.; Majerus, P. W.; Talamo, B.; Vagelos, P. R. Acyl-Carrier Protein. Ii. Intermediary Reactions of Fatty Acid Synthesis. *Biochemistry* **1964**, *3*, 1563−1571.
(19) Majerus, P. W.; Alberts, A. W.; Vagelos, P. R. Acyl Carrier Protein. Iii. An Enoyl Hydrase Specific for Acyl Carrier Protein Thioesters. *J. Biol. Chem.* **1965**, *240*, 618−621.
(20) Weeks, G.; Wakil, S. J. Studies on the Mechanism of Fatty Acid Synthesis. 18. Preparation and General Properties of the Enoyl Acyl Carrier Protein Reductases from Escherichia Coli. *J. Biol. Chem.* **1968**, *243*, 1180−1189.
(21) Roncari, D. A.; Bradshaw, R. A.; Vagelos, P. R. Acyl Carrier Protein. Xix. Amino Acid Sequence of Liver Fatty Acid Synthetase Peptides Containing 4′-Phosphopantetheine. *J. Biol. Chem.* **1972**, *247*, 6234−6242.
(22) Willecke, K.; Ritter, E.; Lynen, F. Isolation of an Acyl Carrier Protein Component from the Multienzyme Complex of Yeast Fatty Acid Synthetase. *Eur. J. Biochem.* **1969**, *8*, 503−509.
(23) Kumar, S.; Dorsey, J. A.; Muesing, R. A.; Porter, J. W. Comparative Studies of the Pigeon Liver Fatty Acid Synthetase Complex and Its Subunits: Kinetics of Partial Reactions and the Number of Binding Sites for Acetyl and Malonyl Groups. *J. Biol. Chem.* **1970**, *245*, 4732−4744.
(24) Kumar, S.; Dorsey, J. K.; Porter, J. W. Mechanism of Dissociation of Pigeon Liver Fatty Acid Synthetase Complex into Half-Molecular Weight Subunits and Their Reassociation to Enzymatically Active Complex. *Biochem. Biophys. Res. Commun.* **1970**, *40*, 825−832.
(25) Smith, S.; Tsai, S.-C. The Type I Fatty Acid and Polyketide Synthases: A Tale of Two Megasynthases. *Nat. Prod. Rep.* **2007**, *24*, 1041−1072.
(26) Rock, C. O.; Jackowski, S. Forty Years of Bacterial Fatty Acid Synthesis. *Biochem. Biophys. Res. Commun.* **2002**, *292*, 1155−1166.
(27) Majerus, P. W. Acyl Carrier Protein. Structural Requirements for Function in Fatty Acid Biosynthesis. *J. Biol. Chem.* **1967**, *242*, 2325−32.
(28) Vanaman, T. C.; Wakil, S. J.; Hill, R. L. The Complete Amino Acid Sequence of the Acyl Carrier Protein of Escherichia Coli. *J. Biol. Chem.* **1968**, *243*, 6420−6431.
(29) Kauppinen, S.; Siggaard-Andersen, M.; von Wettstein-Knowles, P. β -Ketoacyl-Acp Synthase I of Escherichia Coli: Nucleotide Sequence of the Fabb Gene and Identification of the Cerulenin Binding Residue. *Carlsberg Res. Commun.* **1988**, *53*, 357−370.
(30) Cronan, J. E., Jr.; Li, W. B.; Coleman, R.; Narasimhan, M.; de Mendoza, D.; Schwab, J. M. Derived Amino Acid Sequence and Identification of Active Site Residues of *Escherichia Coli* -Hydroxydecanoyl Thioester Dehydrase. *J. Biol. Chem.* **1988**, *263*, 4641−6.
(31) Smith, S.; Witkowski, A.; Joshi, A. K. Structural and Functional Organization of the Animal Fatty Acid Synthase. *Prog. Lipid Res.* **2003**, *42*, 289−317.
(32) McGuire, K. A.; Siggaard-Andersen, M.; Bangera, M. G.; Olsen, J. G.; von Wettstein-Knowles, P. β -Ketoacyl-[Acyl Carrier Protein] Synthase I of Escherichia Coli: Aspects of the Condensation Mechanism Revealed by Analyses of Mutations in the Active Site Pocket. *Biochemistry* **2001**, *40*, 9836−9845.
(33) Huang, W.; Jia, J.; Edwards, P.; Dehesh, K.; Schneider, G.; Lindqvist, Y. Crystal Structure of Beta-Ketoacyl-Acyl Carrier Protein Synthase Ii from E. Coli Reveals the Molecular Architecture of Condensing Enzymes. *EMBO J.* **1998**, *17*, 1183−91.
(34) Qiu, X.; Janson, C. A.; Konstantinidis, A. K.; Nwagwu, S.; Silverman, C.; Smith, W. W.; Khandekar, S.; Lonsdale, J.; Abdel-Meguid, S. S. Crystal Structure of Beta-Ketoacyl-Acyl Carrier Protein Synthase Iii. A Key Condensing Enzyme in Bacterial Fatty Acid Biosynthesis. *J. Biol. Chem.* **1999**, *274*, 36465−71.
(35) Davies, C.; Heath, R. J.; White, S. W.; Rock, C. O. The 1.8 Å Crystal Structure and Active-Site Architecture of t-Ketoacyl-Acyl Carrier Protein Synthase Iii (Fabh) from *Escherichia Coli*. *Structure* **2000**, *8*, 185−95.
(36) Holak, T. A.; Nilges, M.; Prestegard, J. H.; Gronenborn, A. M.; Clore, G. M. Three-Dimensional Structure of Acyl Carrier Protein in Solution Determined by Nuclear Magnetic Resonance and the Combined Use of Dynamical Simulated Annealing and Distance Geometry. *Eur. J. Biochem.* **1988**, *175*, 9−15.
(37) Serre, L.; Verbree, E. C.; Dauter, Z.; Stuitje, A. R.; Derewenda, Z. S. The Escherichia Coli Malonyl-Coa:Acyl Carrier Protein Transacylase at 1.5-a Resolution. Crystal Structure of a Fatty Acid Synthase Component. *J. Biol. Chem.* **1995**, *270*, 12961−4.
(38) Baldock, C.; Rafferty, J. B.; Sedelnikova, S. E.; Baker, P. J.; Stuitje, A. R.; Slabas, A. R.; Hawkes, T. R.; Rice, D. W. A Mechanism of Drug Action Revealed by Structural Studies of Enoyl Reductase. *Science* **1996**, *274*, 2107−2110.
(39) Leesong, M.; Henderson, B. S.; Gillig, J. R.; Schwab, J. M.; Smith, J. L. Structure of a Dehydratase-Isomerase from the Bacterial Pathway for Biosynthesis of Unsaturated Fatty Acids: Two Catalytic Activities in One Active Site. *Structure* **1996**, *4*, 253−264.
(40) Price, A. C.; Zhang, Y.-M.; Rock, C. O.; White, S. W. Structure of β-Ketoacyl-[Acyl Carrier Protein] Reductase from Escherichia Coli: Negative Cooperativity and Its Structural Basis. *Biochemistry* **2001**, *40*, 12772−12781.
(41) Amy, C. M.; Witkowski, A.; Naggert, J.; Williams, B.; Randhawa, Z.; Smith, S. Molecular Cloning and Sequencing of Cdnas Encoding the Entire Rat Fatty Acid Synthase. *Proc. Natl. Acad. Sci. U. S. A.* **1989**, *86*, 3114−8.
(42) Holzer, K. P.; Liu, W.; Hammes, G. G. Molecular Cloning and Sequencing of Chicken Liver Fatty Acid Synthase Cdna. *Proc. Natl. Acad. Sci. U. S. A.* **1989**, *86*, 4387−91.
(43) Schweizer, E.; Kottig, H.; Regler, R.; Rottner, G. Genetic-Control of Yarrowia-Lipolytica Fatty-Acid Synthetase Biosynthesis and Function. *J. Basic Microbiol.* **1988**, *28*, 283−292.
(44) Schweizer, M.; Roberts, L. M.; Höltke, H.-J.; Takabayashi, K.; Höllerer, E.; Hoffmann, B.; Müller, G.; Köttig, H.; Schweizer, E. The Pentafunctional Fas1 Gene of Yeast: Its Nucleotide Sequence and Order of the Catalytic Domains. *Mol. Gen. Genet.* **1986**, *203*, 479−486.
(45) Schweizer, M.; Takabayashi, K.; Laux, T.; Beck, K.-F.; Schreglmann, R. Rat Mammary Gland Fatty Acid Synthase: Localization of the Constituent Domains and Two Functional Polyadenylation/Termination Signals in the Cdna. *Nucleic Acids Res.* **1989**, *17*, 567−586.
(46) Mohamed, A. H.; Chirala, S. S.; Mody, N. H.; Huang, W. Y.; Wakil, S. J. Primary Structure of the Multifunctional Alpha Subunit Protein of Yeast Fatty Acid Synthase Derived from Fas2 Gene Sequence. *J. Biol. Chem.* **1988**, *263*, 12315−25.
(47) Maier, T.; Leibundgut, M.; Boehringer, D.; Ban, N. Structure and Function of Eukaryotic Fatty Acid Synthases. *Q. Rev. Biophys.* **2010**, *43*, 373−422.
(48) Lynen, F. Nobel Lecture - The Pathway from "Activated Acetic Acid" to the Terpenes and Fatty Acids, 1964. https://www.nobelprize.org/prizes/medicine/1964/lynen/lecture/ (accessed 10 Sep 2020).
(49) Pilz, I.; Herbst, M.; Kratky, O.; Oesterhelt, D.; Lynen, F. Röntgenkleinwinkel-Untersuchungen an Der Fettsäuresynthetase Aus Hefe. *Eur. J. Biochem.* **1970**, *13*, 55−64.
(50) Wieland, F.; Siess, E. A.; Renner, L.; Verfürth, C.; Lynen, F. Distribution of Yeast Fatty Acid Synthetase Subunits: Three-Dimensional Model of the Enzyme. *Proc. Natl. Acad. Sci. U. S. A.* **1978**, *75*, 5792−5796.
(51) Stoops, J. K.; Awad, E. S.; Arslanian, M. J.; Gunsberg, S.; Wakil, S. J.; Oliver, R. M. Studies on the Yeast Fatty Acid Synthetase. Subunit Composition and Structural Organization of a Large Multifunctional Enzyme Complex. *J. Biol. Chem.* **1978**, *253*, 4464−75.
(52) Stoops, J. K.; Wakil, S. J. Yeast Fatty Acid Synthetase: Structure-Function Relationship and Nature of the Beta-Ketoacyl Synthetase Site. *Proc. Natl. Acad. Sci. U. S. A.* **1980**, *77*, 4544−4548.
(53) Stoops, J. K.; Wakil, S. J. The Yeast Fatty Acid Synthetase. Structure-Function Relationship and the Role of the Active Cysteine-Sh and Pantetheine-Sh. *J. Biol. Chem.* **1981**, *256*, 8364−70.
(54) Kolodziej, S. J.; Penczek, P. A.; Stoops, J. K. Utility of Butvar Support Film and Methylamine Tungstate Stain in Three-Dimensional Electron Microscopy: Agreement between Stain and Frozen-Hydrated Reconstructions. *J. Struct. Biol.* **1997**, *120*, 158−167.

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

(55) Stoops, J. K.; Kolodziej, S. J.; Schroeter, J. P.; Bretaudiere, J. P.; Wakil, S. J. Structure-Function Relationships of the Yeast Fatty Acid Synthase: Negative-Stain, Cryo-Electron Microscopy, and Image Analysis Studies of the End Views of the Structure. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 6585−6589.
(56) Jenni, S.; Leibundgut, M.; Maier, T.; Ban, N. Architecture of a Fungal Fatty Acid Synthase at 5 Å Resolution. *Science* **2006**, *311*, 1263−1267.
(57) Lomakin, I. B.; Xiong, Y.; Steitz, T. A. The Crystal Structure of Yeast Fatty Acid Synthase, a Cellular Machine with Eight Active Sites Working Together. *Cell* **2007**, *129*, 319−332.
(58) Leibundgut, M.; Jenni, S.; Frick, C.; Ban, N. Structural Basis for Substrate Delivery by Acyl Carrier Protein in the Yeast Fatty Acid Synthase. *Science* **2007**, *316*, 288−290.
(59) Brink, J.; Ludtke, S. J.; Yang, C.-Y.; Gu, Z.-W.; Wakil, S. J.; Chiu, W. Quaternary Structure of Human Fatty Acid Synthase by Electron Cryomicroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 138−143.
(60) Kitamoto, T.; Nishigai, M.; Sasaki, T.; Ikai, A. Structure of Fatty Acid Synthetase from the Harderian Gland of Guinea Pig. Proteolytic Dissection and Electron Microscopic Studies. *J. Mol. Biol.* **1988**, *203*, 183−95.
(61) Stoops, J. K.; Wakil, S. J.; Uberbacher, E. C.; Bunick, G. J. Small-Angle Neutron-Scattering and Electron Microscope Studies of the Chicken Liver Fatty Acid Synthase. *J. Biol. Chem.* **1987**, *262*, 10246−51.
(62) Huang, W. Y.; Chirala, S. S.; Wakil, S. J. Amino-Terminal Blocking Group and Sequence of the Animal Fatty Acid Synthase. *Arch. Biochem. Biophys.* **1994**, *314*, 45−49.
(63) Witkowski, A.; Joshi, A.; Smith, S. Fatty Acid Synthase: In Vitro Complementation of Inactive Mutants. *Biochemistry* **1996**, *35*, 10569−10575.
(64) Witkowski, A.; Rangan, V. S.; Randhawa, Z. I.; Amy, C. M.; Smith, S. Structural Organization of the Multifunctional Animal Fatty-Acid Synthase. *Eur. J. Biochem.* **1991**, *198*, 571−9.
(65) Brink, J.; Ludtke, S. J.; Kong, Y.; Wakil, S. J.; Ma, J.; Chiu, W. Experimental Verification of Conformational Variation of Human Fatty Acid Synthase as Predicted by Normal Mode Analysis. *Structure* **2004**, *12*, 185−191.
(66) Asturias, F. J.; Chadick, J. Z.; Cheung, I. K.; Stark, H.; Witkowski, A.; Joshi, A. K.; Smith, S. Structure and Molecular Organization of Mammalian Fatty Acid Synthase. *Nat. Struct. Mol. Biol.* **2005**, *12*, 225−32.
(67) Oesterhelt, D.; Bauer, H.; Lynen, F. Crystallization of a Multienzyme Complex: Fatty Acid Synthetase from Yeast. *Proc. Natl. Acad. Sci. U. S. A.* **1969**, *63*, 1377−1382.
(68) Maier, T. Architecture of Mammalian Fatty Acid Synthase at 4.5 Å Resolution. *Science* **2006**, *311*, 1258−1262.
(69) Reed, M. A. C.; Schweizer, M.; Szafranska, A. E.; Arthur, C.; Nicholson, T. P.; Cox, R. J.; Crosby, J.; Crump, M. P.; Simpson, T. J. The Type I Rat Fatty Acid Synthase Acp Shows Structural Homology and Analogous Biochemical Properties to Type Ii Acps. *Org. Biomol. Chem.* **2003**, *1*, 463−471.
(70) Chakravarty, B.; Gu, Z.; Chirala, S. S.; Wakil, S. J.; Quiocho, F. A. Human Fatty Acid Synthase: Structure and Substrate Selectivity of the Thioesterase Domain. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 15567−15572.
(71) Greenstein, J. *Biochemistry of Cancer*; 2nd revised ed.; Academic Press: New York, 1954.
(72) Medes, G.; Thomas, A.; Weinhouse, S. Metabolism of Neoplastic Tissue. Iv. A Study of Lipid Synthesis in Neoplastic Tissue Slices in Vitro. *Cancer Res.* **1953**, *13*, 27−29.
(73) Ookhtens, M.; Kannan, R.; Lyon, I.; Baker, N. Liver and Adipose Tissue Contributions to Newly Formed Fatty Acids in an Ascites Tumor. *Am. J. Physiol.* **1984**, *247*, R146−53.
(74) Kuhajda, F. P.; Jenner, K.; Wood, F. D.; Hennigar, R. A.; Jacobs, L. B.; Dick, J. D.; Pasternack, G. R. Fatty Acid Synthesis: A Potential Selective Target for Antineoplastic Therapy. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 6379−83.
(75) Epstein, J. I.; Carmichael, M.; Partin, A. W. Oa-519 (Fatty Acid Synthase) as an Independent Predictor of Pathologic Stage in Adenocarcinoma of the Prostate. *Urology* **1995**, *45*, 81−86.
(76) Shurbaji, M. S.; Kalbfleisch, J. H.; Thurmond, T. S. Immunohistochemical Detection of a Fatty Acid Synthase (Oa-519) as a Predictor of Progression of Prostate Cancer. *Hum. Pathol.* **1996**, *27*, 917−21.
(77) Rashid, A.; Pizer, E. S.; Moga, M.; Milgraum, L. Z.; Zahurak, M.; Pasternack, G. R.; Kuhajda, F. P.; Hamilton, S. R. Elevated Expression of Fatty Acid Synthase and Fatty Acid Synthetic Activity in Colorectal Neoplasia. *Am. J. Pathol.* **1997**, *150*, 201−208.
(78) Gansler, T. S.; Hardman, W., 3rd; Hunt, D. A.; Schaffel, S.; Hennigar, R. A. Increased Expression of Fatty Acid Synthase (Oa-519) in Ovarian Neoplasms Predicts Shorter Survival. *Hum. Pathol.* **1997**, *28*, 686−92.
(79) Pizer, E. S.; Lax, S. F.; Kuhajda, F. P.; Pasternack, G. R.; Kurman, R. J. Fatty Acid Synthase Expression in Endometrial Carcinoma: Correlation with Cell Proliferation and Hormone Receptors. *Cancer* **1998**, *83*, 528−37.
(80) Ito, T.; Sato, K.; Maekawa, H.; Sakurada, M.; Orita, H.; Shimada, K.; Daida, H.; Wada, R.; Abe, M.; Hino, O.; et al. Elevated Levels of Serum Fatty Acid Synthase in Patients with Gastric Carcinoma. *Oncol. Lett.* **2014**, *7*, 616−620.
(81) De Schrijver, E.; Brusselmans, K.; Heyns, W.; Verhoeven, G.; Swinnen, J. V. Rna Interference-Mediated Silencing of the Fatty Acid Synthase Gene Attenuates Growth and Induces Morphological Changes and Apoptosis of Lncap Prostate Cancer Cells. *Cancer Res.* **2003**, *63*, 3799−3804.
(82) Graner, E.; Tang, D.; Rossi, S.; Baron, A.; Migita, T.; Weinstein, L. J; Lechpammer, M.; Huesken, D.; Zimmermann, J.; Signoretti, S.; Loda, M. The Isopeptidase Usp2a Regulates the Stability of Fatty Acid Synthase in Prostate Cancer. *Cancer Cell* **2004**, *5*, 253−261.
(83) Kridel, S. J.; Axelrod, F.; Rozenkrantz, N.; Smith, J. W. Orlistat Is a Novel Inhibitor of Fatty Acid Synthase with Antitumor Activity. *Cancer Res.* **2004**, *64*, 2070−5.
(84) Pizer, E. S.; Thupari, J.; Han, W. F.; Pinn, M. L.; Chrest, F. J.; Frehywot, G. L.; Townsend, C. A.; Kuhajda, F. P. Malonyl-Coenzyme-a Is a Potential Mediator of Cytotoxicity Induced by Fatty-Acid Synthase Inhibition in Human Breast Cancer Cells and Xenografts. *Cancer Res.* **2000**, *60*, 213−218.
(85) Pizer, E. S.; Wood, F. D.; Heine, H. S.; Romantsev, F. E.; Pasternack, G. R.; Kuhajda, F. P. Inhibition of Fatty Acid Synthesis Delays Disease Progression in a Xenograft Model of Ovarian Cancer. *Cancer Res.* **1996**, *56*, 1189−1193.
(86) Royer, G. P.; Townsend, C. A. Cerulenin Compounds for Fatty Acid Synthesis Inhibition. US 5539132, 1996.
(87) Kuhajda, F. P.; Pasternack, G. R. Chemotherapy for Cancer by Inhibiting the Fatty Acid Biosynthetic Pathway. US 5759837, 2 Jun 1998, 1998.
(88) Pizer, E. S.; Kuhajda, F. P.; Townsend, C. A. Cytostatic Effects of Fas Inhibition. US 20020173447, 2002.
(89) Kuhajda, F. P.; Jaffee, E. M.; Townsend, C. A. Method for Inhibiting Cancer Development by Fas Inhibitors, US 20070142456, 21 Jun 2007.
(90) Prochaska, H. J.; Scotto, K. W., Use of Flavonoids to Treat Multidrug Resistant Cancer Cells. US 5336685, 1994.
(91) Chung, F. L. Inhibition of Lung Tumorigenesis by Administration of a Polyphenol. US 5391568, 20 Feb 1995.
(92) Morré, D. M. Tea Catechins as Cancer Specific Proliferation Inhibitors. US 6652890, 2003.
(93) Gorsek, W. F. Cancer Immune Composition for Prevention and Treatment of Individuals, US 20020164317, 7 Nov 2002.
(94) Ceriani, R. L.; Peterson, J. A. Method of Treating Human Breast Cancer by Administration of Radiolabeled Antibody and Unsaturated Fatty Acids. US 5766571, 16 June 1998.
(95) Obukowicz, M. G. Combinations of Omega-3 Fatty Acids and Cyclooxygenase-2 Inhibitors for Treatment or Prevention of Cardiovascular Disease and Treatment of Cancer. US 20040044028, 2004.

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

(96) Pichette, A.; Legault, J.; Madelmont, J. Antitumor Methods and Compositions Comprising Sesquiterpene Derivatives. US 20090197949, 6 Aug 2009.
(97) Smith, J. W., Axelrod, F., Kridel, S. J., Inhibition of Fas by β-Lactones and Other Compounds for Inhibition of Cellular Proliferation. US 7799826, 2010.
(98) Loftus, T. M.; Jaworsky, D. E.; Frehywot, G. L.; Townsend, C. A.; Ronnett, G. V.; Lane, M. D.; Kuhajda, F. P. Reduced Food Intake and Body Weight in Mice Treated with Fatty Acid Synthase Inhibitors. *Science* **2000**, *288*, 2379−2381.
(99) Mobbs, C. V.; Makimura, H. Block the Fas, Lose the Fat. *Nat. Med.* **2002**, *8*, 335−6.
(100) Berndt, J.; Kovacs, P.; Ruschke, K.; Klöting, N.; Fasshauer, M.; Schön, M. R.; Körner, A.; Stumvoll, M.; Blüher, M. Fatty Acid Synthase Gene Expression in Human Adipose Tissue: Association with Obesity and Type 2 Diabetes. *Diabetologia* **2007**, *50*, 1472−80.
(101) Junien, J.; Edgar, A. Use of Metformin and Orlistat for the Treatment or Prevention of Obesity. US 20070060532, 15 Mar 2007.
(102) Edgar, A. Use of Fibrate and Orlistat for the Treatment of Obesity. US 20070078179, 5 Apr 2007.
(103) Menendez, J. A.; Vellon, L.; Lupu, R. Antitumoral Actions of the Anti-Obesity Drug Orlistat (Xenicaltm) in Breast Cancer Cells: Blockade of Cell Cycle Progression, Promotion of Apoptotic Cell Death and Pea3-Mediated Transcriptional Repression of Her2/Neu (Erbb-2) Oncogene. *Ann. Oncol.* **2005**, *16*, 1253−67.
(104) Rittner, A.; Paithankar, K. S.; Huu, K. V.; Grininger, M. Characterization of the Polyspecific Transferase of Murine Type I Fatty Acid Synthase (Fas) and Implications for Polyketide Synthase (Pks) Engineering. *ACS Chem. Biol.* **2018**, *13*, 723−732.
(105) Khosla, C.; Kapur, S.; Cane, D. E. Revisiting the Modularity of Modular Polyketide Synthases. *Curr. Opin. Chem. Biol.* **2009**, *13*, 135−143.
(106) Keatinge-Clay, A. T. The Structures of Type I Polyketide Synthases. *Nat. Prod. Rep.* **2012**, *29*, 1050−1073.
(107) Katsuyama, Y.; Ohnishi, Y. Type Iii Polyketide Synthases in Microorganisms. *Methods Enzymol.* **2012**, *515*, 359−377.
(108) Maier, T.; Leibundgut, M.; Ban, N. The Crystal Structure of a Mammalian Fatty Acid Synthase. *Science* **2008**, *321*, 1315−1322.
(109) Keatinge-Clay, A. T.; Stroud, R. M. The Structure of a Ketoreductase Determines the Organization of the Beta-Carbon Processing Enzymes of Modular Polyketide Synthases. *Structure* **2006**, *14*, 737−748.
(110) Jenke-Kodama, H.; Sandmann, A.; Muller, R.; Dittmann, E. Evolutionary Implications of Bacterial Polyketide Synthases. *Mol. Biol. Evol.* **2005**, *22*, 2027−2039.
(111) Rittner, A.; Paithankar, K. S.; Drexler, D. J.; Himmler, A.; Grininger, M. Probing the Modularity of Megasynthases by Rational Engineering of a Fatty Acid Synthase Type I. *Protein Sci.* **2019**, *28*, 414−428.
(112) Zhang, W.; Chakravarty, B.; Zheng, F.; Gu, Z.; Wu, H.; Mao, J.; Wakil, S. J.; Quiocho, F. A. Crystal Structure of Fas Thioesterase Domain with Polyunsaturated Fatty Acyl Adduct and Inhibition by Dihomo-Γ-Linolenic Acid. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 15757−15762.
(113) Jayakumar, A.; Huang, W. Y.; Raetz, B.; Chirala, S. S.; Wakil, S. J. Cloning and Expression of the Multifunctional Human Fatty Acid Synthase and Its Subdomains in Escherichia Coli. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 14509−14514.
(114) Stuible, H. P.; Meier, S.; Schweizer, E. Identification, Isolation and Biochemical Characterization of a Phosphopantetheine:Protein Transferase That Activates the Two Type-I Fatty Acid Synthases of Brevibacterium Ammoniagenes. *Eur. J. Biochem.* **1997**, *248*, 481−487.
(115) Ciccarelli, L.; Connell, S. R.; Enderle, M.; Mills, D. J.; Vonck, J.; Grininger, M. Structure and Conformational Variability of the *Mycobacterium Tuberculosis* Fatty Acid Synthase Multienzyme Complex. *Structure* **2013**, *21*, 1251−1257.
(116) Leber, C.; Da Silva, N. A. Engineering of Saccharomyces Cerevisiae for the Synthesis of Short Chain Fatty Acids. *Biotechnol. Bioeng.* **2014**, *111*, 347−358.
(117) Witkowski, A.; Ghosal, A.; Joshi, A. K.; Witkowska, H. E.; Asturias, F. J.; Smith, S. Head-to-Head Coiled Arrangement of the Subunits of the Animal Fatty Acid Synthase. *Chem. Biol.* **2004**, *11*, 1667−76.
(118) Viegas, M. F.; Neves, R. P. P.; Ramos, M. J.; Fernandes, P. A. Modeling of Human Fatty Acid Synthase and in Silico Docking of Acyl Carrier Protein Domain and Its Partner Catalytic Domains. *J. Phys. Chem. B* **2018**, *122*, 77−85.
(119) Smith, S.; Abraham, S. Reversible Dissociation of Fatty Acid Synthetase from Lactating Rat Mammary Gland. *Biochem. J.* **1970**, *120*, 31P−32P.
(120) Stoops, J. K.; Wakil, S. J. Animal Fatty Acid Synthetase. A Novel Arrangement of the Beta-Ketoacyl Synthetase Sites Comprising Domains of the Two Subunits. *J. Biol. Chem.* **1981**, *256*, 5128−33.
(121) Stoops, J. K.; Wakil, S. J. Animal Fatty-Acid Synthetase - Identification of the Residues Comprising the Novel Arrangement of the Beta-Ketoacyl Synthetase Site and Their Role in Its Cold Inactivation. *J. Biol. Chem.* **1982**, *257*, 3230−3235.
(122) Wakil, S. J. Fatty Acid Synthase, a Proficient Multifunctional Enzyme. *Biochemistry* **1989**, *28*, 4523−30.
(123) Joshi, A. K.; Witkowski, A.; Smith, S. The Malonyl/Acetyltransferase and Beta-Ketoacyl Synthase Domains of the Animal Fatty Acid Synthase Can Cooperate with the Acyl Carrier Protein Domain of Either Subunit. *Biochemistry* **1998**, *37*, 2515−2523.
(124) Swarbrick, C. M. D.; Nanson, J. D.; Patterson, E. I.; Forwood, J. K. Structure, Function, and Regulation of Thioesterases. *Prog. Lipid Res.* **2020**, *79*, 101036.
(125) Gago, G.; Diacovich, L.; Arabolaza, A.; Tsai, S. C.; Gramajo, H. Fatty Acid Biosynthesis in Actinomycetes. *FEMS Microbiol. Rev.* **2011**, *35*, 475−497.
(126) Jenni, S.; Leibundgut, M.; Boehringer, D.; Frick, C.; Mikolásek, B.; Ban, N. Structure of Fungal Fatty Acid Synthase and Implications for Iterative Substrate Shuttling. *Science* **2007**, *316*, 254−261.
(127) Johansson, P.; Mulinacci, B.; Koestler, C.; Vollrath, R.; Oesterhelt, D.; Grininger, M. Multimeric Options for the Auto-Activation of the Saccharomyces Cerevisiae Fas Type I Megasynthase. *Structure* **2009**, *17*, 1063−1074.
(128) Singh, K.; Graf, B.; Linden, A.; Sautner, V.; Urlaub, H.; Tittmann, K.; Stark, H.; Chari, A. Discovery of a Regulatory Subunit of the Yeast Fatty Acid Synthase. *Cell* **2020**, *180*, 1130−1143.
(129) Anselmi, C.; Grininger, M.; Gipson, P.; Faraldo-Gomez, J. D. Mechanism of Substrate Shuttling by the Acyl-Carrier Protein within the Fatty Acid Mega-Synthase. *J. Am. Chem. Soc.* **2010**, *132*, 12357−12364.
(130) Boehringer, D.; Ban, N.; Leibundgut, M. 7.5-Å Cryo-Em Structure of the Mycobacterial Fatty Acid Synthase. *J. Mol. Biol.* **2013**, *425*, 841−849.
(131) Joshi, A. K.; Witkowski, A.; Berman, H. A.; Zhang, L.; Smith, S. Effect of Modification of the Length and Flexibility of the Acyl Carrier Protein-Thioesterase Interdomain Linker on Functionality of the Animal Fatty Acid Synthase. *Biochemistry* **2005**, *44*, 4100−4107.
(132) Brignole, E. J.; Smith, S.; Asturias, F. J. Conformational Flexibility of Metazoan Fatty Acid Synthase Enables Catalysis. *Nat. Struct. Mol. Biol.* **2009**, *16*, 190−197.
(133) Benning, F. M. C.; Sakiyama, Y.; Mazur, A.; Bukhari, H. S. T.; Lim, R. Y. H.; Maier, T. High-Speed Atomic Force Microscopy Visualization of the Dynamics of the Multienzyme Fatty Acid Synthase. *ACS Nano* **2017**, *11*, 10852−10859.
(134) Jensen-Urstad, A. P. L.; Semenkovich, C. F. Fatty Acid Synthase and Liver Triglyceride Metabolism: Housekeeper or Messenger? *Biochim. Biophys. Acta, Mol. Cell Biol. Lipids* **2012**, *1821*, 747−753.
(135) Lehner, R.; Quiroga, A. D. In *Biochemistry of Lipids, Lipoproteins and Membranes*, 6th ed.; Ridgway, N. D., McLeod, R. S., Eds.; Elsevier: Boston, 2016.
(136) Foretz, M.; Guichard, C.; Ferré, P.; Foufelle, F. Sterol Regulatory Element Binding Protein-1c Is a Major Mediator of Insulin Action on the Hepatic Expression of Glucokinase and Lipogenesis-Related Genes. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 12737−12742.

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

(137) Ishii, S.; Iizuka, K.; Miller, B. C.; Uyeda, K. Carbohydrate Response Element Binding Protein Directly Promotes Lipogenic Enzyme Gene Transcription. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 15597−15602.
(138) Moon, Y. S.; Latasa, M.-J.; Griffin, M. J.; Sul, H. S. Suppression of Fatty Acid Synthase Promoter by Polyunsaturated Fatty Acids. *J. Lipid Res.* **2002**, *43*, 691−698.
(139) Dentin, R.; Benhamed, F.; Pégorier, J.-P.; Foufelle, F.; Viollet, B.; Vaulont, S.; Girard, J.; Postic, C. Polyunsaturated Fatty Acids Suppress Glycolytic and Lipogenic Genes through the Inhibition of Chrebp Nuclear Protein Translocation. *J. Clin. Invest.* **2005**, *115*, 2843−2854.
(140) Horton, J. D.; Bashmakov, Y.; Shimomura, I.; Shimano, H. Regulation of Sterol Regulatory Element Binding Proteins in Livers of Fasted and Refed Mice. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 5987−5992.
(141) Najjar, S. M.; Yang, Y.; Fernstrom, M. A.; Lee, S.-J.; DeAngelis, A. M.; Rjaily, G. A. A.; Al-Share, Q. Y.; Dai, T.; Miller, T. A.; Ratnam, S.; Ruch, R. J.; Smith, S.; Lin, S.-H.; Beauchemin, N.; Oyarce, A. M. Insulin Acutely Decreases Hepatic Fatty Acid Synthase Activity. *Cell Metab.* **2005**, *2*, 43−53.
(142) Qureshi, A. A.; Jenik, R. A.; Kim, M.; Lornitzo, F. A.; Porter, J. W. Separation of Two Active Forms (Holo-a and Holo-B) of Pigeon Liver Fatty Acid Synthetase and Their Interconversion by Phosphorylation and Dephosphorylation. *Biochem. Biophys. Res. Commun.* **1975**, *66*, 344−351.
(143) Zhao, S.; Xu, W.; Jiang, W.; Yu, W.; Lin, Y.; Zhang, T.; Yao, J.; Zhou, L.; Zeng, Y.; Li, H.; et al. Regulation of Cellular Metabolism by Protein Lysine Acetylation. *Science* **2010**, *327*, 1000−1004.
(144) Chakravarthy, M. V.; Lodhi, I. J.; Yin, L.; Malapaka, R. R.; Xu, H. E.; Turk, J.; Semenkovich, C. F. Identification of a Physiologically Relevant Endogenous Ligand for Pparα in Liver. *Cell* **2009**, *138*, 476−488.
(145) Motojima, K.; Passilly, P.; Peters, J. M.; Gonzalez, F. J.; Latruffe, N. Expression of Putative Fatty Acid Transporter Genes Are Regulated by Peroxisome Proliferator-Activated Receptor A and Γ Activators in a Tissue-and Inducer-Specific Manner. *J. Biol. Chem.* **1998**, *273*, 16710−16714.
(146) Leone, T. C.; Weinheimer, C. J.; Kelly, D. P. A Critical Role for the Peroxisome Proliferator-Activated Receptor A (Pparα) in the Cellular Fasting Response: The Pparα-Null Mouse as a Model of Fatty Acid Oxidation Disorders. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 7473−7478.
(147) Dreyer, C.; Keller, H.; Mahfoudi, A.; Laudet, V.; Krey, G.; Wahli, W. Positive Regulation of the Peroxisomal β-Oxidation Pathway by Fatty Acids through Activation of Peroxisome Proliferator-Activated Receptors (Ppar). *Biol. Cell* **1993**, *77*, 67−74.
(148) Aoyama, T.; Peters, J. M.; Iritani, N.; Nakajima, T.; Furihata, K.; Hashimoto, T.; Gonzalez, F. J. Altered Constitutive Expression of Fatty Acid-Metabolizing Enzymes in Mice Lacking the Peroxisome Proliferator-Activated Receptor A (Pparα). *J. Biol. Chem.* **1998**, *273*, 5678−5684.
(149) Wu, M.; Singh, S. B.; Wang, J.; Chung, C. C.; Salituro, G.; Karanam, B. V.; Lee, S. H.; Powles, M.; Ellsworth, K. P.; Lassman, M. E.; et al. Antidiabetic and Antisteatotic Effects of the Selective Fatty Acid Synthase (Fas) Inhibitor Platensimycin in Mouse Models of Diabetes. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 5378−5383.
(150) Weiss, L.; Hoffmann, G. E.; Schreiber, R.; Andres, H.; Fuchs, E.; Körber, E.; Kolb, H. J. Fatty-Acid Biosynthesis in Man, a Pathway of Minor Importance. Purification, Optimal Assay Conditions, and Organ Distribution of Fatty-Acid Synthase. *Biol. Chem. Hoppe-Seyler* **1986**, *367*, 905−12.
(151) Vazquez-Martin, A.; Colomer, R.; Brunet, J.; Lupu, R.; Menendez, J. A. Overexpression of Fatty Acid Synthase Gene Activates Her1/Her2 Tyrosine Kinase Receptors in Human Breast Epithelial Cells. *Cell Proliferation* **2008**, *41*, 59−85.
(152) Gonzalez-Guerrico, A. M.; Espinoza, I.; Schroeder, B.; Park, C. H.; Kvp, C. M.; Khurana, A.; Corominas-Faja, B.; Cuyàs, E.; Alarcón, T.; Kleer, C.; et al. Suppression of Endogenous Lipogenesis Induces Reversion of the Malignant Phenotype and Normalized Differentiation in Breast Cancer. *Oncotarget* **2016**, *7*, 71151−71168.
(153) Fiorentino, M.; Zadra, G.; Palescandolo, E.; Fedele, G.; Bailey, D.; Fiore, C.; Nguyen, P. L.; Migita, T.; Zamponi, R.; Di Vizio, D.; et al. Overexpression of Fatty Acid Synthase Is Associated with Palmitoylation of Wnt1 and Cytoplasmic Stabilization of β-Catenin in Prostate Cancer. *Lab. Invest.* **2008**, *88*, 1340−1348.
(154) Rossi, S.; Ou, W.; Tang, D.; Bhattacharya, N.; Dei Tos, A. P.; Fletcher, J. A.; Loda, M. Gastrointestinal Stromal Tumours Overexpress Fatty Acid Synthase. *J. Pathol.* **2006**, *209*, 369−75.
(155) Wang, W.-q.; Zhao, X.-y.; Wang, H.-y.; Liang, Y. Increased Fatty Acid Synthase as a Potential Therapeutic Target in Multiple Myeloma. *J. Zhejiang Univ., Sci., B* **2008**, *9*, 441−447.
(156) Little, J. L.; Kridel, S. J. In *Lipids in Health and Disease*; Quinn, P. J., Wang, X., Eds.; Springer Netherlands: Dordrecht, 2008. DOI: 10.1007/978-1-4020-8831-5_7.
(157) Wu, X.; Qin, L.; Fako, V.; Zhang, J. T. Molecular Mechanisms of Fatty Acid Synthase (Fasn)-Mediated Resistance to Anti-Cancer Treatments. *Adv. Biol. Regul.* **2014**, *54*, 214−21.
(158) Zadra, G.; Photopoulos, C.; Loda, M. The Fat Side of Prostate Cancer. *Biochim. Biophys. Acta, Mol. Cell Biol. Lipids* **2013**, *1831*, 1518−32.
(159) Shah, U. S.; Dhir, R.; Gollin, S. M.; Chandran, U. R.; Lewis, D.; Acquafondata, M.; Pflug, B. R. Fatty Acid Synthase Gene Overexpression and Copy Number Gain in Prostate Adenocarcinoma. *Hum. Pathol.* **2006**, *37*, 401−9.
(160) Ogino, S.; Nosho, K.; Meyerhardt, J. A.; Kirkner, G. J.; Chan, A. T.; Kawasaki, T.; Giovannucci, E. L.; Loda, M.; Fuchs, C. S. Cohort Study of Fatty Acid Synthase Expression and Patient Survival in Colon Cancer. *J. Clin. Oncol.* **2008**, *26*, 5713−5720.
(161) Alo, P. L.; Visca, P.; Marci, A.; Mangoni, A.; Botti, C.; Di Tondo, U. Expression of Fatty Acid Synthase (Fas) as a Predictor of Recurrence in Stage I Breast Carcinoma Patients. *Cancer* **1996**, *77*, 474−82.
(162) Kao, Y. C.; Lee, S. W.; Lin, L. C.; Chen, L. T.; Hsing, C. H.; Hsu, H. P.; Huang, H. Y.; Shiue, Y. L.; Chen, T. J.; Li, C. F. Fatty Acid Synthase Overexpression Confers an Independent Prognosticator and Associates with Radiation Resistance in Nasopharyngeal Carcinoma. *Tumor Biol.* **2013**, *34*, 759−68.
(163) Duan, J.; Sun, L.; Huang, H.; Wu, Z.; Wang, L.; Liao, W. Overexpression of Fatty Acid Synthase Predicts a Poor Prognosis for Human Gastric Cancer. *Mol. Med. Rep.* **2016**, *13*, 3027−35.
(164) Cheng, J.; Ondracek, R. P.; Mehedint, D. C.; Kasza, K. A.; Xu, B.; Gill, S.; Azabdaftari, G.; Yao, S.; Morrison, C. D.; Mohler, J. L.; et al. Association of Fatty-Acid Synthase Polymorphisms and Expression with Outcomes after Radical Prostatectomy. *Prostate Cancer Prostatic Dis.* **2015**, *18*, 182−189.
(165) Buckley, D.; Duke, G.; Heuer, T. S.; O'Farrell, M.; Wagman, A. S.; McCulloch, W.; Kemble, G. Fatty Acid Synthase - Modern Tumor Cell Biology Insights into a Classical Oncology Target. *Pharmacol. Ther.* **2017**, *177*, 23−31.
(166) Funabashi, H.; Kawaguchi, A.; Tomoda, H.; Omura, S.; Okuda, S.; Iwasaki, S. Binding Site of Cerulenin in Fatty Acid Synthetase. *J. Biochem.* **1989**, *105*, 751−5.
(167) Yang, P.-Y.; Liu, K.; Ngai, M. H.; Lear, M. J.; Wenk, M. R.; Yao, S. Q. Activity-Based Proteome Profiling of Potential Cellular Targets of Orlistat - an Fda-Approved Drug with Anti-Tumor Activities. *J. Am. Chem. Soc.* **2010**, *132*, 656−666.
(168) Brusselmans, K.; De Schrijver, E.; Heyns, W.; Verhoeven, G.; Swinnen, J. V. Epigallocatechin-3-Gallate Is a Potent Natural Inhibitor of Fatty Acid Synthase in Intact Cells and Selectively Induces Apoptosis in Prostate Cancer Cells. *Int. J. Cancer* **2003**, *106*, 856−862.
(169) Hardwicke, M. A.; Rendina, A. R.; Williams, S. P.; Moore, M. L.; Wang, L.; Krueger, J. A.; Plant, R. N.; Totoritis, R. D.; Zhang, G.; Briand, J.; et al. A Human Fatty Acid Synthase Inhibitor Binds Beta-Ketoacyl Reductase in the Keto-Substrate Site. *Nat. Chem. Biol.* **2014**, *10*, 774−779.
(170) Johansson, P.; Wiltschi, B.; Kumari, P.; Kessler, B.; Vonrhein, C.; Vonck, J.; Oesterhelt, D.; Grininger, M. Inhibition of the Fungal

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

Fatty Acid Synthase Type I Multienzyme Complex. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 12803−12808.
(171) Rendina, A. R.; Cheng, D. Characterization of the Inactivation of Rat Fatty Acid Synthase by C75: Inhibition of Partial Reactions and Protection by Substrates. *Biochem. J.* **2005**, *388*, 895−903.
(172) Pizer, E. S.; Jackisch, C.; Wood, F. D.; Pasternack, G. R.; Davidson, N. E.; Kuhajda, F. P. Inhibition of Fatty Acid Synthesis Induces Programmed Cell Death in Human Breast Cancer Cells. *Cancer Res.* **1996**, *56*, 2745−2747.
(173) Alli, P. M.; Pinn, M. L.; Jaffee, E. M.; McFadden, J. M.; Kuhajda, F. P. Fatty Acid Synthase Inhibitors Are Chemopreventive for Mammary Cancer in Neu-N Transgenic Mice. *Oncogene* **2005**, *24*, 39−46.
(174) Flavin, R.; Peluso, S.; Nguyen, P. L.; Loda, M. Fatty Acid Synthase as a Potential Therapeutic Target in Cancer. *Future Oncol.* **2010**, *6*, 551−62.
(175) Pemble, C. W.; Johnson, L. C.; Kridel, S. J.; Lowther, W. T. Crystal Structure of the Thioesterase Domain of Human Fatty Acid Synthase Inhibited by Orlistat. *Nat. Struct. Mol. Biol.* **2007**, *14*, 704−709.
(176) Jones, S. F.; Infante, J. R. Molecular Pathways: Fatty Acid Synthase. *Clin. Cancer Res.* **2015**, *21*, 5434−8.
(177) Chen, W. L.; Chen, Y. L.; Chiang, Y. M.; Wang, S. G.; Lee, H. M. Fenofibrate Lowers Lipid Accumulation in Myotubes by Modulating the Pparα/Ampk/Foxo1/Atgl Pathway. *Biochem. Pharmacol.* **2012**, *84*, 522−31.
(178) You, B.-J.; Hour, M.-J.; Chen, L.-Y.; Luo, S.-C.; Hsu, P.-H.; Lee, H.-Z. Fenofibrate Induces Human Hepatoma Hep3b Cells Apoptosis and Necroptosis through Inhibition of Thioesterase Domain of Fatty Acid Synthase. *Sci. Rep.* **2019**, *9*, 3306.
(179) Alwarawrah, Y.; Hughes, P.; Loiselle, D.; Carlson, D. A.; Darr, D. B.; Jordan, J. L.; Xiong, J.; Hunter, L. M.; Dubois, L. G.; Thompson, J. W.; et al. Fasnall, a Selective Fasn Inhibitor, Shows Potent Anti-Tumor Activity in the Mmtv-Neu Model of Her2(+) Breast Cancer. *Cell Chem. Biol.* **2016**, *23*, 678−88.
(180) Angeles, T. S.; Hudkins, R. L. Recent Advances in Targeting the Fatty Acid Biosynthetic Pathway Using Fatty Acid Synthase Inhibitors. *Expert Opin. Drug Discovery* **2016**, *11*, 1187−1199.
(181) Mullen, G. E.; Yet, L. Progress in the Development of Fatty Acid Synthase Inhibitors as Anticancer Targets. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 4363−4369.
(182) Chang, S. I.; Hammes, G. G. Structure and Mechanism of Action of a Multifunctional Enzyme: Fatty Acid Synthase. *Acc. Chem. Res.* **1990**, *23*, 363−369.
(183) Chirala, S. S.; Wakil, S. J. Structure and Function of Animal Fatty Acid Synthase. *Lipids* **2004**, *39*, 1045−53.
(184) Smith, S. The Animal Fatty Acid Synthase: One Gene, One Polypeptide, Seven Enzymes. *FASEB J.* **1994**, *8*, 1248−59.
(185) Fischbach, M. A.; Walsh, C. T. Assembly-Line Enzymology for Polyketide and Nonribosomal Peptide Antibiotics: Logic, Machinery, and Mechanisms. *Chem. Rev.* **2006**, *106*, 3468−96.
(186) Khosla, C.; Tang, Y.; Chen, A. Y.; Schnarr, N. A.; Cane, D. E. Structure and Mechanism of the 6-Deoxyerythronolide B Synthase. *Annu. Rev. Biochem.* **2007**, *76*, 195−221.
(187) McCarthy, A. D.; Aitken, A.; Hardie, D. G.; Santikarn, S.; Williams, D. H. Amino Acid Sequence around the Active Serine in the Acyl Transferase Domain of Rabbit Mammary Fatty Acid Synthase. *FEBS Lett.* **1983**, *160*, 296−300.
(188) McCarthy, A. D.; Hardie, D. G. The Multifunctional Polypeptide Chains of Rabbit-Mammary Fatty-Acid Synthase. Stoichiometry of Active Sites and Active-Site Mapping Using Limited Proteolysis. *Eur. J. Biochem.* **1983**, *130*, 185−93.
(189) Schweizer, E.; Lerch, I.; Kroeplin-Rueff, L.; Lynen, F. Fatty Acyl Transferase: Characterization of the Enzyme as Part of the Yeast Fatty Acid Synthetase Complex by the Use of Radioactively Labeled Coenzyme A. *Eur. J. Biochem.* **1970**, *15*, 472−82.
(190) Ayling, J.; Pirson, R.; Lynen, F. Participation of Covalently Linked Fatty Acyl Coenzyme A Products in the Action of Yeast Fatty Acid Synthetase. *Biochemistry* **1972**, *11*, 526−33.
(191) Engeser, H.; HüBner, K.; Straub, J.; Lynen, F. Identity of Malonyl and Palmitoyl Transferase of Fatty Acid Synthetase from Yeast. *Eur. J. Biochem.* **1979**, *101*, 407−412.
(192) Heath, R. J.; Rock, C. O. Regulation of Fatty Acid Elongation and Initiation by Acyl-Acyl Carrier Protein in Escherichia Coli. *J. Biol. Chem.* **1996**, *271*, 1833−6.
(193) Joshi, V. C.; Wakil, S. J. Studies on the Mechanism of Fatty Acid Synthesis. Xxvi. Purification and Properties of Malonyl-Coenzyme A–Acyl Carrier Protein Transacylase of Escherichia Coli. *Arch. Biochem. Biophys.* **1971**, *143*, 493−505.
(194) LYNEN, F.; Engeser, H.; Foerster, E.-C.; Fox, J. L.; Hess, S.; Kresze, G.-B.; Schmitt, T.; Schreckenbach, T.; Siess, E.; Wieland, F.; Winnewisser, W. On the Structure of Fatty Acid Synthetase of Yeast. *Eur. J. Biochem.* **1980**, *112*, 431−442.
(195) Schuster, H.; Rautenstrauss, B.; Mittag, M.; Stratmann, D.; Schweizer, E. Substrate and Product Binding Sites of Yeast Fatty Acid Synthase. *Eur. J. Biochem.* **1995**, *228*, 417−424.
(196) White, S. W.; Zheng, J.; Zhang, Y.-M.; Rock, C. O. The Structural Biology of Type Ii Fatty Acid Biosynthesis. *Annu. Rev. Biochem.* **2005**, *74*, 791−831.
(197) Williamson, I. P.; Wakil, S. J. Studies on the Mechanism of Fatty Acid Synthesis. Xvii. Preparation and General Properties of Acetyl Coenzyme A and Malonyl Coenzyme A-Acyl Carrier Protein Transacylases. *J. Biol. Chem.* **1966**, *241*, 2326−32.
(198) Rangan, V. S.; Witkowski, A.; Smith, S. Isolation of a Functional Transferase Component from the Rat Fatty Acid Synthase by Limited Trypsinization of the Subunit Monomer. Formation of a Stable Functional Complex between Transferase and Acyl Carrier Protein Domains. *J. Biol. Chem.* **1991**, *266*, 19180−5.
(199) Stern, A.; Sedgwick, B.; Smith, S. The Free Coenzyme A Requirement of Animal Fatty Acid Synthetase. Participation in the Continuous Exchange of Acetyl and Malonyl Moieties between Coenzyme A Thioester and Enzyme. *J. Biol. Chem.* **1982**, *257*, 799−803.
(200) Joshi, V. C.; Plate, C. A.; Wakil, S. J. Studies on the Mechanism of Fatty Acid Synthesis. 23. The Acyl Binding Sites of the Pigeon Liver Fatty Acid Synthetase. *J. Biol. Chem.* **1970**, *245*, 2857−67.
(201) Mikkelsen, J.; Hojrup, P.; Rasmussen, M. M.; Roepstorff, P.; Knudsen, J. Amino Acid Sequence around the Active-Site Serine Residue in the Acyltransferase Domain of Goat Mammary Fatty Acid Synthetase. *Biochem. J.* **1985**, *227*, 21−7.
(202) Phillips, G. T.; Nixon, J. E.; Dorsey, J. A.; Butterworth, P. H.; Chesterton, C. J.; Porter, J. W. The Mechanism of Synthesis of Fatty Acids by the Pigeon Liver Enzyme System. *Arch. Biochem. Biophys.* **1970**, *138*, 380−91.
(203) Plate, C. A.; Joshi, V. C.; Wakil, S. J. Studies on the Mechanism of Fatty Acid Synthesis. Xxiv. The Acetyl- and Malonyltransacylase Activities of Pigeon Liver Fatty Acid Synthetase. *J. Biol. Chem.* **1970**, *245*, 2868−2875.
(204) Soulie, J. M.; Sheplock, G. J.; Tian, W. X.; Hsu, R. Y. Transient Kinetic Studies of Fatty Acid Synthetase. A Kinetic Self-Editing Mechanism for the Loading of Acetyl and Malonyl Residues and the Role of Coenzyme A. *J. Biol. Chem.* **1984**, *259*, 134−40.
(205) Buckner, J. S.; Kolattukudy, P. E.; Rogers, L. Synthesis of Multimethyl-Branched Fatty Acids by Avian and Mammalian Fatty Acid Synthetase and Its Regulation by Malonyl-Coa Decarboxylase in the Uropygial Gland. *Arch. Biochem. Biophys.* **1978**, *186*, 152−63.
(206) Heil, C. S.; Wehrheim, S. S.; Paithankar, K. S.; Grininger, M. Fatty Acid Biosynthesis: Chain-Length Regulation and Control. *ChemBioChem* **2019**, *20*, 2298−2321.
(207) Seyama, Y.; Otsuka, H.; Kawaguchi, A.; Yamakawa, T. Fatty Acid Synthetase from the Harderian Gland of Guinea Pig: Biosynthesis of Methyl-Branched Fatty Acids. *J. Biochem.* **1981**, *90*, 789−97.
(208) Smith, S.; Stern, A. The Effect of Aromatic Coa Esters on Fatty Acid Synthetase: Biosynthesis of Omega-Phenyl Fatty Acids. *Arch. Biochem. Biophys.* **1983**, *222*, 259−65.
(209) Wallace, M.; Green, C. R.; Roberts, L. S.; Lee, Y. M.; McCarville, J. L.; Sanchez-Gurmaches, J.; Meurs, N.; Gengatharan, J. M.; Hover, J. D.; Phillips, S. A.; et al. Enzyme Promiscuity Drives

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

Branched-Chain Fatty Acid Synthesis in Adipose Tissues. *Nat. Chem. Biol.* **2018**, *14*, 1021−1031.
(210) Schweizer, E.; Piccinini, F.; Duba, C.; GÜNther, S.; Ritter, E.; Lynen, F. Die Malonyl-Bindungsstellen Des Fettsäuresynthetase-Komplexes Aus Hefe. *Eur. J. Biochem.* **1970**, *15*, 483−499.
(211) Cleland, W. W. The Kinetics of Enzyme-Catalyzed Reactions with Two or More Substrates or Products: I. Nomenclature and Rate Equations. *Biochim. Biophys. Acta, Spec. Sect. Enzymol. Subj.* **1963**, *67*, 104−137.
(212) Dunn, B. J.; Cane, D. E.; Khosla, C. Mechanism and Specificity of an Acyltransferase Domain from a Modular Polyketide Synthase. *Biochemistry* **2013**, *52*, 1839−1841.
(213) Lowe, P. N.; Rhodes, S. Purification and Characterization of [Acyl-Carrier-Protein] Acetyltransferase from Escherichia Coli. *Biochem. J.* **1988**, *250*, 789−96.
(214) Ruch, F. E.; Vagelos, P. R. The Isolation and General Properties of Escherichia Coli Malonyl Coenzyme A-Acyl Carrier Protein Transacylase. *J. Biol. Chem.* **1973**, *248*, 8086−8094.
(215) Szafranska, A. E.; Hitchman, T. S.; Cox, R. J.; Crosby, J.; Simpson, T. J. Kinetic and Mechanistic Analysis of the Malonyl Coa:Acp Transacylase from Streptomyces Coelicolor Indicates a Single Catalytically Competent Serine Nucleophile at the Active Site. *Biochemistry* **2002**, *41*, 1421−1427.
(216) Holmquist, M. Alpha/Beta-Hydrolase Fold Enzymes: Structures, Functions and Mechanisms. *Curr. Protein Pept. Sci.* **2000**, *1*, 209−35.
(217) Nardini, M.; Dijkstra, B. W. α/β Hydrolase Fold Enzymes: The Family Keeps Growing. *Curr. Opin. Struct. Biol.* **1999**, *9*, 732−737.
(218) Ollis, D. L.; Cheah, E.; Cygler, M.; Dijkstra, B.; Frolow, F.; Franken, S. M.; Harel, M.; Remington, S. J.; Silman, I.; Schrag, J.; et al. The Alpha/Beta Hydrolase Fold. *Protein Eng., Des. Sel.* **1992**, *5*, 197−211.
(219) Rangan, V. S.; Smith, S. Expression in Escherichia Coli and Refolding of the Malonyl-/Acetyltransferase Domain of the Multifunctional Animal Fatty Acid Synthase. *J. Biol. Chem.* **1996**, *271*, 31749−55.
(220) Oefner, C.; Schulz, H.; D'Arcy, A.; Dale, G. E. Mapping the Active Site of Escherichia Coli Malonyl-Coa-Acyl Carrier Protein Transacylase (Fabd) by Protein Crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, *62*, 613−8.
(221) Bunkoczi, G.; Misquitta, S.; Wu, X.; Lee, W. H.; Rojkova, A.; Kochan, G.; Kavanagh, K. L.; Oppermann, U.; Smith, S. Structural Basis for Different Specificities of Acyltransferases Associated with the Human Cytosolic and Mitochondrial Fatty Acid Synthases. *Chem. Biol.* **2009**, *16*, 667−75.
(222) Gokhale, R. S.; Lau, J.; Cane, D. E.; Khosla, C. Functional Orientation of the Acyltransferase Domain in a Module of the Erythromycin Polyketide Synthase. *Biochemistry* **1998**, *37*, 2524−2528.
(223) Paiva, P.; Sousa, S. F.; Ramos, M. J.; Fernandes, P. A. Understanding the Catalytic Machinery and the Reaction Pathway of the Malonyl-Acetyl Transferase Domain of Human Fatty Acid Synthase. *ACS Catal.* **2018**, *8*, 4860−4872.
(224) Rangan, V. S.; Smith, S. Alteration of the Substrate Specificity of the Malonyl-Coa/Acetyl-Coa:Acyl Carrier Protein S-Acyltransferase Domain of the Multifunctional Fatty Acid Synthase by Mutation of a Single Arginine Residue. *J. Biol. Chem.* **1997**, *272*, 11975−8.
(225) Liew, C. W.; Nilsson, M.; Chen, M. W.; Sun, H.; Cornvik, T.; Liang, Z. X.; Lescar, J. Crystal Structure of the Acyltransferase Domain of the Iterative Polyketide Synthase in Enediyne Biosynthesis. *J. Biol. Chem.* **2012**, *287*, 23203−15.
(226) Paiva, P.; Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Human Fatty Acid Synthase: A Computational Study of the Transfer of the Acyl Moieties from Mat to the Acp Domain. *ChemCatChem* **2019**, *11*, 3853−3864.
(227) Laidler, K. J.; King, M. C. Development of Transition-State Theory. *J. Phys. Chem.* **1983**, *87*, 2657−2664.
(228) Keatinge-Clay, A. T.; Shelat, A. A.; Savage, D. F.; Tsai, S. C.; Miercke, L. J.; O'Connell, J. D., 3rd; Khosla, C.; Stroud, R. M. Catalysis, Specificity, and Acp Docking Site of Streptomyces Coelicolor Malonyl-Coa:Acp Transacylase. *Structure* **2003**, *11*, 147−54.
(229) Bryan, P.; Pantoliano, M. W.; Quill, S. G.; Hsiao, H. Y.; Poulos, T. Site-Directed Mutagenesis and the Role of the Oxyanion Hole in Subtilisin. *Proc. Natl. Acad. Sci. U. S. A.* **1986**, *83*, 3743−5.
(230) Kraut, D. A.; Sigala, P. A.; Pybus, B.; Liu, C. W.; Ringe, D.; Petsko, G. A.; Herschlag, D. Testing Electrostatic Complementarity in Enzyme Catalysis: Hydrogen Bonding in the Ketosteroid Isomerase Oxyanion Hole. *PLoS Biol.* **2006**, *4*, No. e99.
(231) Warshel, A.; Russell, S. Theoretical Correlation of Structure and Energetics in the Catalytic Reaction of Trypsin. *J. Am. Chem. Soc.* **1986**, *108*, 6569−6579.
(232) Heath, R. J.; Rock, C. O. The Claisen Condensation in Biology. *Nat. Prod. Rep.* **2002**, *19*, 581−596.
(233) Khosla, C.; Gokhale, R. S.; Jacobsen, J. R.; Cane, D. E. Tolerance and Specificity of Polyketide Synthases. *Annu. Rev. Biochem.* **1999**, *68*, 219−53.
(234) Herbst, D. A.; Townsend, C. A.; Maier, T. The Architectures of Iterative Type I Pks and Fas. *Nat. Prod. Rep.* **2018**, *35*, 1046−1069.
(235) Witkowski, A.; Joshi, A. K.; Smith, S. Characterization of the Interthiol Acyltransferase Reaction Catalyzed by the Beta-Ketoacyl Synthase Domain of the Animal Fatty Acid Synthase. *Biochemistry* **1997**, *36*, 16338−44.
(236) D'Agnolo, G.; Rosenfeld, I. S.; Vagelos, P. R. Multiple Forms of (-Ketoacyl-Acyl Carrier Protein Synthetase in *Escherichia Coli*. *J. Biol. Chem.* **1975**, *250*, 5289−94.
(237) Garwin, J. L.; Klages, A. L.; Cronan, J. E., Jr. Structural, Enzymatic, and Genetic Studies of Beta-Ketoacyl-Acyl Carrier Protein Synthases I and Ii of Escherichia Coli. *J. Biol. Chem.* **1980**, *255*, 11949−56.
(238) Rosenfeld, I. S.; D'Agnolo, G.; Vagelos, P. R. Synthesis of Unsaturated Fatty Acids and the Lesion in Fab B Mutants. *J. Biol. Chem.* **1973**, *248*, 2452−60.
(239) Tsay, J. T.; Oh, W.; Larson, T. J.; Jackowski, S.; Rock, C. O. Isolation and Characterization of the Beta-Ketoacyl-Acyl Carrier Protein Synthase Iii Gene (Fabh) from Escherichia Coli K-12. *J. Biol. Chem.* **1992**, *267*, 6807−14.
(240) Siggaard-Andersen, M. Conserved Residues in Condensing Enzyme Domains of Fatty Acid Synthases and Related Sequences. *Protein Seq. Data Anal.* **1993**, *5*, 325−335.
(241) Christensen, C. E.; Kragelund, B. B.; von Wettstein-Knowles, P.; Henriksen, A. Structure of the Human (-Ketoacyl [Acp] Synthase from the Mitochondrial Type Ii Fatty Acid Synthase. *Protein Sci.* **2006**, *16*, 261−72.
(242) Moche, M.; Dehesh, K.; Edwards, P.; Lindqvist, Y. The Crystal Structure of Beta-Ketoacyl-Acyl Carrier Protein Synthase Ii from Synechocystis Sp. At 1.54 a Resolution and Its Relationship to Other Condensing Enzymes. *J. Mol. Biol.* **2001**, *305*, 491−503.
(243) Olsen, J. G.; Kadziola, A.; von Wettstein-Knowles, P.; Siggaard-Andersen, M.; Larsen, S. Structures of Beta-Ketoacyl-Acyl Carrier Protein Synthase I Complexed with Fatty Acids Elucidate Its Catalytic Machinery. *Structure* **2001**, *9*, 233−43.
(244) Pan, H.; Tsai, S.; Meadows, E. S.; Miercke, L. J.; Keatinge-Clay, A. T.; O'Connell, J.; Khosla, C.; Stroud, R. M. Crystal Structure of the Priming Beta-Ketosynthase from the R1128 Polyketide Biosynthetic Pathway. *Structure* **2002**, *10*, 1559−68.
(245) Pappenberger, G.; Benz, J.; Gsell, B.; Hennig, M.; Ruf, A.; Stihle, M.; Thoma, R.; Rudolph, M. G. Structure of the Human Fatty Acid Synthase Ks-Mat Didomain as a Framework for Inhibitor Design. *J. Mol. Biol.* **2010**, *397*, 508−19.
(246) Qiu, X.; Janson, C. A.; Smith, W. W.; Head, M.; Lonsdale, J.; Konstantinidis, A. K. Refined Structures of Beta-Ketoacyl-Acyl Carrier Protein Synthase Iii. *J. Mol. Biol.* **2001**, *307*, 341−56.
(247) Tang, Y. Y.; Kim, C. Y.; Mathews, I. I.; Cane, D. E.; Khosla, C. The 2.7-Angstrom Crystal Structure of a 194-Kda Homodimeric Fragment of the 6-Deoxyerythronolide B Synthase. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 11124−11129.
(248) Rittner, A.; Paithankar, K. S.; Himmler, A.; Grininger, M. Type I Fatty Acid Synthase Trapped in the Octanoyl-Bound State. *Protein Sci.* **2020**, *29*, 589−605.

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

(249) Hol, W. G. J.; van Duijnen, P. T.; Berendsen, H. J. C. The A-Helix Dipole and the Properties of Proteins. *Nature* **1978**, *273*, 443−446.
(250) Joshi, A. K.; Witkowski, A.; Smith, S. Mapping of Functional Interactions between Domains of the Animal Fatty Acid Synthase by Mutant Complementation in Vitro. *Biochemistry* **1997**, *36*, 2316−2322.
(251) Witkowski, A.; Joshi, A. K.; Lindqvist, Y.; Smith, S. Conversion of a Beta-Ketoacyl Synthase to a Malonyl Decarboxylase by Replacement of the Active-Site Cysteine with Glutamine. *Biochemistry* **1999**, *38*, 11643−50.
(252) Witkowski, A.; Joshi, A. K.; Smith, S. Mechanism of the Beta-Ketoacyl Synthase Reaction Catalyzed by the Animal Fatty Acid Synthase. *Biochemistry* **2002**, *41*, 10877−87.
(253) Zhang, Y. M.; Hurlbert, J.; White, S. W.; Rock, C. O. Roles of the Active Site Water, Histidine 303, and Phenylalanine 396 in the Catalytic Mechanism of the Elongation Condensing Enzyme of Streptococcus Pneumoniae. *J. Biol. Chem.* **2006**, *281*, 17390−9.
(254) Bressler, R.; Wakil, S. J. Studies on the Mechanism of Fatty Acid Synthesis: Ix. The Conversion of Malonyl Coenzyme A to Long Chain Fatty Acids. *J. Biol. Chem.* **1961**, *236*, 1643−1651.
(255) Wakil, S. J.; Stoops, J. K.; Joshi, V. C. Fatty Acid Synthesis and Its Regulation. *Annu. Rev. Biochem.* **1983**, *52*, 537−579.
(256) Witkowski, A.; Joshi, A. K.; Rangan, V. S.; Falick, A. M.; Witkowska, H. E.; Smith, S. Dibromopropanone Cross-Linking of the Phosphopantetheine and Active-Site Cysteine Thiols of the Animal Fatty Acid Synthase Can Occur Both Inter- and Intrasubunit. *J. Biol. Chem.* **1999**, *274*, 11557−11563.
(257) Witkowski, A.; Joshi, A. K.; Smith, S. Characterization of the Beta-Carbon Processing Reactions of the Mammalian Cytosolic Fatty Acid Synthase: Role of the Central Core. *Biochemistry* **2004**, *43*, 10458−10466.
(258) Porter, J. W.; Wakil, S. J.; Tietz, A.; Jacob, M. I.; Gibson, D. M. Studies on the Mechanism of Fatty Acid Synthesis. Ii. Cofactor Requirements of the Soluble Pigeon Liver System. *Biochim. Biophys. Acta* **1957**, *25*, 35−41.
(259) Wakil, S. J.; Porter, J. W.; Gibson, D. M. Studies on the Mechanism of Fatty Acid Synthesis. I. Preparation and Purification of an Enzymes System for Reconstruction of Fatty Acid Synthesis. *Biochim. Biophys. Acta* **1957**, *24*, 453−461.
(260) Brady, R. O. Biosynthesis of Fatty Acids Ii. Studies with Enzymes Obtained from Brain. *J. Biol. Chem.* **1960**, *235*, 3099−3103.
(261) Bressler, R.; Wakil, S. J. Studies on the Mechanism of Fatty Acid Synthesis. Xi. The Product of the Reaction and the Role of Sulfhydryl Groups in the Synthesis of Fatty Acids. *J. Biol. Chem.* **1962**, *237*, 1441−8.
(262) Alberts, A. W.; Goldman, P.; Vagelos, P. R. The Condensation Reaction of Fatty Acid Aynthesis. I. Separation and Properties of the Enzymes. *J. Biol. Chem.* **1963**, *238*, 557−65.
(263) Lynen, F. In *Methods in Enzymology*; Academic Press, 1969; Vol. *14*.
(264) Alberts, A. W.; Majerus, P. W.; Vagelos, P. R. Acyl Carrier Protein. Vi. Purification and Properties of *β*-Ketoacyl Acyl Carrier Protein Synthetase*. *Biochemistry* **1965**, *4*, 2265−2274.
(265) Toomey, R. E.; Wakil, S. J. Studies on the Mechanism of Fatty Acid Synthesis. Xvi. Preparation and General Properties of Acyl-Malonyl Acyl Carrier Protein-Condensing Enzyme from Escherichia Coli. *J. Biol. Chem.* **1966**, *241*, 1159−65.
(266) Greenspan, M. D.; Alberts, A. W.; Vagelos, P. R. Acyl Carrier Protein. 13. Beta-Ketoacyl Acyl Carrier Protein Synthetase from Escherichia Coli. *J. Biol. Chem.* **1969**, *244*, 6477−85.
(267) Lynen, F. The Role of Biotin-Dependent Carboxylations in Biosynthetic Reactions. *Biochem. J.* **1967**, *102*, 381−400.
(268) Kresze, G. B.; Steber, L.; Oesterhelt, D.; Lynen, F. Reaction of Yeast Fatty Acid Synthetase with Iodoacetamide. 3. Malonyl-Coenzyme A Decarboxylase as Product of the Reaction of Fatty Acid Synthetase with Iodoacetamide. *Eur. J. Biochem.* **1977**, *79*, 191−199.
(269) Oesterhelt, D.; Bauer, H.; Kresze, G. B.; Steber, L.; Lynen, F. Reaction of Yeast Fatty Acid Synthetase with Iodoacetamide. 1. Kinetics of Inactivation and Extent of Carboxamidomethylation. *Eur. J. Biochem.* **1977**, *79*, 173−180.
(270) Jez, J. M.; Noel, J. P. Mechanism of Chalcone Synthase: Pka of the Catalytic Cysteine and the Role of the Conserved Histidine in a Plant Polyketide Synthase. *J. Biol. Chem.* **2000**, *275*, 39640−39646.
(271) Anand, S.; Mohanty, D. Modeling Holo-Acp:Dh and Holo-Acp:Kr Complexes of Modular Polyketide Synthases: A Docking and Molecular Dynamics Study. *BMC Struct. Biol.* **2012**, *12*, 10.
(272) Hu, B.; Guo, H.; Zhou, P.; Shi, Z. L. Characteristics of Sars-Cov-2 and Covid-19. *Nat. Rev. Microbiol.* **2021**, *19*, 141−154.
(273) Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; et al. Structure of M(Pro) from Sars-Cov-2 and Discovery of Its Inhibitors. *Nature* **2020**, *582*, 289−293.
(274) Wu, F.; Zhao, S.; Yu, B.; Chen, Y. M.; Wang, W.; Song, Z. G.; Hu, Y.; Tao, Z. W.; Tian, J. H.; Pei, Y. Y.; et al. A New Coronavirus Associated with Human Respiratory Disease in China. *Nature* **2020**, *579*, 265−269.
(275) Lee, W.; Engels, B. The Protonation State of Catalytic Residues in the Resting State of Kasa Revisited: Detailed Mechanism for the Activation of Kasa by Its Own Substrate. *Biochemistry* **2014**, *53*, 919−31.
(276) Lee, W.; Engels, B. Clarification on the Decarboxylation Mechanism in Kasa Based on the Protonation State of Key Residues in the Acyl-Enzyme State. *J. Phys. Chem. B* **2013**, *117*, 8095−104.
(277) Dreier, J.; Khosla, C. Mechanistic Analysis of a Type Ii Polyketide Synthase. Role of Conserved Residues in the R-Ketoacyl Synthase-Chain Length Factor Heterodimer. *Biochemistry* **2000**, *39*, 2088−95.
(278) Price, A. C.; Choi, K. H.; Heath, R. J.; Li, Z.; White, S. W.; Rock, C. O. Inhibition of (-Ketoacyl-Acyl Carrier Protein Synthases by Thiolactomycin and Cerulenin. Structure and Mechanism. *J. Biol. Chem.* **2001**, *276*, 6551−9.
(279) Sumper, M.; Oesterhelt, D.; Riepertinger, C.; Lynen, F. Die Synthese Verschiedener Carbonsäuren Durch Den Multienzymkomplex Der Fettsäuresynthese Aus Hefe Und Die Erklärung Ihrer Bildung. *Eur. J. Biochem.* **1969**, *10*, 377−387.
(280) Nguyen, T.; Ishida, K.; Jenke-Kodama, H.; Dittmann, E.; Gurgui, C.; Hochmuth, T.; Taudien, S.; Platzer, M.; Hertweck, C.; Piel, J. Exploiting the Mosaic Structure of Trans-Acyltransferase Polyketide Synthases for Natural Product Discovery and Pathway Dissection. *Nat. Biotechnol.* **2008**, *26*, 225−33.
(281) Rangan, V. S.; Joshi, A. K.; Smith, S. Mapping the Functional Topology of the Animal Fatty Acid Synthase by Mutant Complementation in Vitro. *Biochemistry* **2001**, *40*, 10792−10799.
(282) John, A.; Umashankar, V.; Krishnakumar, S.; Deepa, P. R. Comparative Modeling and Molecular Dynamics Simulation of Substrate Binding in Human Fatty Acid Synthase: Enoyl Reductase and *β*-Ketoacyl Reductase Catalytic Domains. *Genomics Inform.* **2015**, *13*, 15−24.
(283) Persson, B.; Kallberg, Y.; Oppermann, U.; Jornvall, H. Coenzyme-Based Functional Assignments of Short-Chain Dehydrogenases/Reductases (Sdrs). *Chem.-Biol. Interact.* **2003**, *143*, 271−278.
(284) Balakrishnan, A.; Gao, Y.; Moorjani, P.; Nemeria, N. S.; Tittmann, K.; Jordan, F. Bifunctionality of the Thiamin Diphosphate Cofactor: Assignment of Tautomeric/Ionization States of the 4'-Aminopyrimidine Ring When Various Intermediates Occupy the Active Sites During the Catalysis of Yeast Pyruvate Decarboxylase. *J. Am. Chem. Soc.* **2012**, *134*, 3873−3885.
(285) Cohen-Gonsaud, M.; Ducasse, S.; Hoh, F.; Zerbib, D.; Labesse, G.; Quemard, A. k. Crystal Structure of Maba from *Mycobacterium Tuberculosis*, a Reductase Involved in Long-Chain Fatty Acid Biosynthesis. *J. Mol. Biol.* **2002**, *320*, 249−261.
(286) Oppermann, U.; Filling, C.; Hult, M.; Shafqat, N.; Wu, X.; Lindh, M.; Shafqat, J.; Nordling, E.; Kallberg, Y.; Persson, B.; et al. Short-Chain Dehydrogenases/Reductases (Sdr): The 2002 Update. *Chem.-Biol. Interact.* **2003**, *143−144*, 247−253.
(287) Jörnvall, H. Medium- and Short-Chain Dehydrogenase/Reductase Gene and Protein Families: Mdr and Sdr Gene and Protein Superfamilies. *Cell. Mol. Life Sci.* **2008**, *65*, 3873−3878.

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

(288) Dutta, D.; Bhattacharyya, S.; Roychowdhury, A.; Biswas, R.; Das, A. K. Crystal Structure of Hexanoyl-Coa Bound to (-Ketoacyl Reductase Fabg4 of *Mycobacterium Tuberculosis*. *Biochem. J.* **2013**, *450*, 127−139.
(289) Keatinge-Clay, A. T. A Tylosin Ketoreductase Reveals How Chirality Is Determined in Polyketides. *Chem. Biol.* **2007**, *14*, 898−908.
(290) Price, A. C.; Zhang, Y. M.; Rock, C. O.; White, S. W. Cofactor-Induced Conformational Rearrangements Establish a Catalytically Competent Active Site and a Proton Relay Conduit in Fabg. *Structure* **2004**, *12*, 417−428.
(291) Skiba, M. A.; Sikkema, A. P.; Fiers, W. D.; Gerwick, W. H.; Sherman, D. H.; Aldrich, C. C.; Smith, J. L. Domain Organization and Active Site Architecture of a Polyketide Synthase C-Methyltransferase. *ACS Chem. Biol.* **2016**, *11*, 3319−3327.
(292) Martin, M. W.; Lancia, D. R.; Li, H.; Schiller, S. E. R.; Toms, A. V.; Wang, Z.; Bair, K. W.; Castro, J.; Fessler, S.; Gotur, D.; et al. Discovery and Optimization of Novel Piperazines as Potent Inhibitors of Fatty Acid Synthase (Fasn). *Bioorg. Med. Chem. Lett.* **2019**, *29*, 1001−1006.
(293) Medina, F. E.; Neves, R. P. P.; Ramos, M. J.; Fernandes, P. A. A Qm/Mm Study of the Reaction Mechanism of Human Beta-Ketoacyl Reductase. *Phys. Chem. Chem. Phys.* **2017**, *19*, 347−355.
(294) Zheng, J.; Keatinge-Clay, A. T. Structural and Functional Analysis of C2-Type Ketoreductases from Modular Polyketide Synthases. *J. Mol. Biol.* **2011**, *410*, 105−17.
(295) Carlisle-Moore, L.; Gordon, C. R.; Machutta, C. A.; Miller, W. T.; Tonge, P. J. Substrate Recognition by the Human Fatty-Acid Synthase. *J. Biol. Chem.* **2005**, *280*, 42612−42618.
(296) Siskos, A. P.; Baerga-Ortiz, A.; Bali, S.; Stein, V.; Mamdani, H.; Spiteller, D.; Popovic, B.; Spencer, J. B.; Staunton, J.; Weissman, K. J.; et al. Molecular Basis of Celmer's Rules: Stereochemistry of Catalysis by Isolated Ketoreductase Domains from Modular Polyketide Synthases. *Chem. Biol.* **2005**, *12*, 1145−53.
(297) Xie, X.; Garg, A.; Keatinge-Clay, A. T.; Khosla, C.; Cane, D. E. Epimerase and Reductase Activities of Polyketide Synthase Ketoreductase Domains Utilize the Same Conserved Tyrosine and Serine Residues. *Biochemistry* **2016**, *55*, 1179−1186.
(298) Bailey, C. B.; Pasman, M. E.; Keatinge-Clay, A. T. Substrate Structure-Activity Relationships Guide Rational Engineering of Modular Polyketide Synthase Ketoreductases. *Chem. Commun.* **2016**, *52*, 792−795.
(299) Baerga-Ortiz, A.; Popovic, B.; Siskos, A. P.; O'Hare, H. M.; Spiteller, D.; Williams, M. G.; Campillo, N.; Spencer, J. B.; Leadlay, P. F. Directed Mutagenesis Alters the Stereochemistry of Catalysis by Isolated Ketoreductase Domains from the Erythromycin Polyketide Synthase. *Chem. Biol.* **2006**, *13*, 277−85.
(300) Mugnai, M. L.; Shi, Y.; Keatinge-Clay, A. T.; Elber, R. Molecular Dynamics Studies of Modular Polyketide Synthase Ketoreductase Stereospecificity. *Biochemistry* **2015**, *54*, 2346−2359.
(301) Fuxreiter, M.; Warshel, A. Origin of the Catalytic Power of Acetylcholinesterase: Computer Simulation Studies. *J. Am. Chem. Soc.* **1998**, *120*, 183−194.
(302) Warshel, A.; Sharma, P. K.; Chu, Z. T.; Aqvist, J. Electrostatic Contributions to Binding of Transition State Analogues Can Be Very Different from the Corresponding Contributions to Catalysis: Phenolates Binding to the Oxyanion Hole of Ketosteroid Isomerase. *Biochemistry* **2007**, *46*, 1466−1476.
(303) Zhang, Y. K.; Kua, J.; McCammon, J. A. Role of the Catalytic Triad and Oxyanion Hole in Acetylcholinesterase Catalysis: An Ab Initio Qm/Mm Study. *J. Am. Chem. Soc.* **2002**, *124*, 10572−10577.
(304) Warshel, A.; Narayszabo, G.; Sussman, F.; Hwang, J. K. How Do Serine Proteases Really Work. *Biochemistry* **1989**, *28*, 3629−3637.
(305) Warshel, A.; Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol.* **1976**, *103*, 227−249.
(306) Senn, H. M.; Thiel, W. Qm/Mm Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198−1229.
(307) Sousa, S. F.; Ribeiro, A. J. M.; Neves, R. P. P.; Brás, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J. Application of Quantum Mechanics/Molecular Mechanics Methods in the Study of Enzymatic Reaction Mechanisms. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2017**, *7*, No. e1281.
(308) Ehrensberger, A. H.; Wilson, D. K. Structural and Catalytic Diversity in the Two Family 11 Aldo-Keto Reductases. *J. Mol. Biol.* **2004**, *337*, 661−673.
(309) Świderek, K.; Kohen, A.; Moliner, V. The Influence of Active Site Conformations on the Hydride Transfer Step of the Thymidylate Synthase Reaction Mechanism. *Phys. Chem. Chem. Phys.* **2015**, *17*, 30793−30804.
(310) Kwan, D. H.; Schulz, F. The Stereochemistry of Complex Polyketide Biosynthesis by Modular Polyketide Synthases. *Molecules* **2011**, *16*, 6092−6115.
(311) Wakil, S. J. Mechanism of Fatty Acid Synthesis*. *J. Lipid Res.* **1961**, *2*, 1−24.
(312) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L. Y.; He, J. E.; He, S. Q.; Shoemaker, B. A.; et al. Pubchem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.
(313) Dillon, S. C.; Bateman, A. The Hotdog Fold: Wrapping up a Superfamily of Thioesterases and Dehydratases. *BMC Bioinf.* **2004**, *5*, 109.
(314) Labonte, J. W.; Townsend, C. A. Active Site Comparisons and Catalytic Mechanisms of the Hot Dog Superfamily. *Chem. Rev.* **2013**, *113*, 2182−2204.
(315) Zhang, W.; Chakravarty, B.; Zheng, F.; Gu, Z. W.; Wu, H. M.; Mao, J. Q.; Wakil, S. J.; Quiocho, F. A. Crystal Structure of Fas Thioesterase Domain with Polyunsaturated Fatty Acyl Adduct and Inhibition by Dihomo-Gamma-Linolenic Acid. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 15757−15762.
(316) Joshi, A. K.; Smith, S. Construction, Expression, and Characterization of a Mutated Animal Fatty-Acid Synthase Deficient in the Dehydratase Function. *J. Biol. Chem.* **1993**, *268*, 22508−22513.
(317) Pasta, S.; Witkowski, A.; Joshi, A. K.; Smith, S. Catalytic Residues Are Shared between Two Pseudosubunits of the Dehydratase Domain of the Animal Fatty Acid Synthase. *Chem. Biol.* **2007**, *14*, 1377−1385.
(318) Li, Y.; Dodge, G. J.; Fiers, W. D.; Fecik, R. A.; Smith, J. L.; Aldrich, C. C. Functional Characterization of a Dehydratase Domain from the Pikromycin Polyketide Synthase. *J. Am. Chem. Soc.* **2015**, *137*, 7003−7006.
(319) Biswas, R.; Dutta, A.; Dutta, D.; Hazra, D.; Banerjee, D. R.; Basak, A.; Das, A. K. Crystal Structure of Dehydratase Component Hadab Complex of Mycobacterial Fas-Ii Pathway. *Biochem. Biophys. Res. Commun.* **2015**, *458*, 369−374.
(320) Medina, F. E.; Neves, R. P. P.; Ramos, M. J.; Fernandes, P. A. Qm/Mm Study of the Reaction Mechanism of the Dehydratase Domain from Mammalian Fatty Acid Synthase. *ACS Catal.* **2018**, *8*, 10267−10278.
(321) Annand, R. R.; Kozlowski, J. F.; Davisson, V. J.; Schwab, J. M. Mechanism-Based Inactivation of *Escherichia Coli*. Beta-Hydroxydecanoyl Thiol Ester Dehydrase: Assignment of the Imidazole Nitrogen-15 Nmr Resonances and Determination of the Structure of the Alkylated Histidine. *J. Am. Chem. Soc.* **1993**, *115*, 1088−1094.
(322) Xie, X.; Cane, D. E. Ph-Rate Profiles Establish That Polyketide Synthase Dehydratase Domains Utilize a Single-Base Mechanism. *Org. Biomol. Chem.* **2018**, *16*, 9165−9170.
(323) Liu, L.; Yu, Q.; Zhang, H.; Tao, W.; Wang, R.; Bai, L.; Zhao, Y.-L.; Shi, T. Theoretical Study on Substrate Recognition and Catalytic Mechanisms of Gephyronic Acid Dehydratase Dh1. *Catal. Sci. Technol.* **2021**, *11*, 2155−2166.
(324) Bunkoczi, G.; Pasta, S.; Joshi, A.; Wu, X.; Kavanagh, K. L.; Smith, S.; Oppermann, U. Mechanism and Substrate Recognition of Human Holo Acp Synthase. *Chem. Biol.* **2007**, *14*, 1243−1253.
(325) Evans, C. T.; Kurz, L. C.; Remington, S. J.; Srere, P. A. Active Site Mutants of Pig Citrate Synthase: Effects of Mutations on the

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

Enzyme Catalytic and Structural Properties. *Biochemistry* **1996**, *35*, 10661−72.
(326) Hamed, R. B.; Batchelar, E. T.; Clifton, I. J.; Schofield, C. J. Mechanisms and Structures of Crotonase Superfamily Enzymes - How Nature Controls Enolate and Oxyanion Reactivity. *Cell. Mol. Life Sci.* **2008**, *65*, 2507−2527.
(327) Shafqat, N.; Shafqat, J.; Eissner, G.; Marschall, H. U.; Tryggvason, K.; Eriksson, U.; Gabrielli, F.; Lardy, H.; Jornvall, H.; Oppermann, U. Hep27, a Member of the Short-Chain Dehydrogenase/Reductase Family, Is an Nadph-Dependent Dicarbonyl Reductase Expressed in Vascular Endothelial Tissue. *Cell. Mol. Life Sci.* **2006**, *63*, 1205−1213.
(328) Rozwarski, D. A.; Vilcheze, C.; Sugantino, M.; Bittman, R.; Sacchettini, J. C. Crystal Structure of the Mycobacterium Tuberculosis Enoyl-Acp Reductase, Inha, in Complex with Nad(+) and a C16 Fatty Acyl Substrate. *J. Biol. Chem.* **1999**, *274*, 15582−15589.
(329) Sippel, K. H.; Vyas, N. K.; Zhang, W.; Sankaran, B.; Quiocho, F. A. Crystal Structure of the Human Fatty Acid Synthase Enoyl-Acyl Carrier Protein-Reductase Domain Complexed with Triclosan Reveals Allosteric Protein-Protein Interface Inhibition. *J. Biol. Chem.* **2014**, *289*, 33287−33295.
(330) Khare, D.; Hale, W. A.; Tripathi, A.; Gu, L.; Sherman, D. H.; Gerwick, W. H.; Håkansson, K.; Smith, J. L. Structural Basis for Cyclopropanation by a Unique Enoyl-Acyl Carrier Protein Reductase. *Structure* **2015**, *23*, 2213−2223.
(331) Massengo-Tiassé, R. P.; Cronan, J. E. Diversity in Enoyl-Acyl Carrier Protein Reductases. *Cell. Mol. Life Sci.* **2009**, *66*, 1507−1517.
(332) Toogood, H. S.; Scrutton, N. S. Discovery, Characterization, Engineering, and Applications of Ene-Reductases for Industrial Biocatalysis. *ACS Catal.* **2018**, *8*, 3532−3549.
(333) Melo, A. M. P.; Bandeiras, T. M.; Teixeira, M. New Insights into Type Ii Nad(P)H:Quinone Oxidoreductases. *Microbiol. Mol. Biol. Rev.* **2004**, *68*, 603.
(334) Sena, F. V.; Sousa, F. M.; Oliveira, A. S. F.; Soares, C. M.; Catarino, T.; Pereira, M. M. Regulation of the Mechanism of Type-Ii Nadh: Quinone Oxidoreductase from S. Aureus. *Redox Biol.* **2018**, *16*, 209−214.
(335) Hedlund, J.; Jörnvall, H.; Persson, B. Subdivision of the Mdr Superfamily of Medium-Chain Dehydrogenases/Reductases through Iterative Hidden Markov Model Refinement. *BMC Bioinf.* **2010**, *11*, 534.
(336) Nordling, E.; Jörnvall, H.; Persson, B. Medium-Chain Dehydrogenases/Reductases (Mdr). Family Characterizations Including Genome Comparisons and Active Site Modeling. *Eur. J. Biochem.* **2002**, *269*, 4267−76.
(337) Persson, B.; Hedlund, J.; Jörnvall, H. Medium- and Short-Chain Dehydrogenase/Reductase Gene and Protein Families. *Cell. Mol. Life Sci.* **2008**, *65*, 3879.
(338) Erb, T. J.; Berg, I. A.; Brecht, V.; Müller, M.; Fuchs, G.; Alber, B. E. Synthesis of CS-Dicarboxylic Acids from C2-Units Involving Crotonyl-Coa Carboxylase/Reductase: The Ethylmalonyl-Coa Pathway. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 10631.
(339) Quade, N.; Huo, L.; Rachid, S.; Heinz, D. W.; Müller, R. Unusual Carbon Fixation Gives Rise to Diverse Polyketide Extender Units. *Nat. Chem. Biol.* **2012**, *8*, 117−124.
(340) Heath, R. J.; Rock, C. O. Enoyl-Acyl Carrier Protein Reductase (Fabi) Plays a Determinant Role in Completing Cycles of Fatty-Acid Elongation in E. Coli. *J. Biol. Chem.* **1995**, *270*, 26538−26542.
(341) Rosenthal, R. G.; Ebert, M.-O.; Kiefer, P.; Peter, D. M.; Vorholt, J. A.; Erb, T. J. Direct Evidence for a Covalent Ene Adduct Intermediate in Nad(P)H-Dependent Enzymes. *Nat. Chem. Biol.* **2014**, *10*, 50−55.
(342) Rosenthal, R. G.; Vogeli, B.; Quade, N.; Capitani, G.; Kiefer, P.; Vorholt, J. A.; Ebert, M. O.; Erb, T. J. The Use of Ene Adducts to Study and Engineer Enoyl-Thioester Reductases. *Nat. Chem. Biol.* **2015**, *11*, 398−400.
(343) Kwan, D. H.; Leadlay, P. F. Mutagenesis of a Modular Polyketide Synthase Enoylreductase Domain Reveals Insights into Catalysis and Stereospecificity. *ACS Chem. Biol.* **2010**, *5*, 829−838.
(344) Rosenthal, R. G.; Vogeli, B.; Wagner, T.; Shima, S.; Erb, T. J. A Conserved Threonine Prevents Self-Intoxication of Enoyl-Thioester Reductases. *Nat. Chem. Biol.* **2017**, *13*, 745−749.
(345) Vogeli, B.; Erb, T. J. 'Negative' and 'Positive Catalysis': Complementary Principles That Shape the Catalytic Landscape of Enzymes. *Curr. Opin. Chem. Biol.* **2018**, *47*, 94−100.
(346) Tallorin, L.; Finzel, K.; Nguyen, Q. G.; Beld, J.; La Clair, J. J.; Burkart, M. D. Trapping of the Enoyl-Acyl Carrier Protein Reductase-Acyl Carrier Protein Interaction. *J. Am. Chem. Soc.* **2016**, *138*, 3962−3965.
(347) Abeles, R. H.; Hutton, R. F.; Westheimer, F. H. The Reduction of Thioketones by a Model for a Coenzyme. *J. Am. Chem. Soc.* **1957**, *79*, 712−716.
(348) You, K.-S.; Oppenheimer, N. J. Stereospecificity for Nicotinamide Nucleotides in Enzymatic and Chemical Hydride Transfer-Reactions. *Crit. Rev. Biochem. Mol. Biol.* **1985**, *17*, 313−451.
(349) Medina, F. E.; Ramos, M. J.; Fernandes, P. A. Complexities of the Reaction Mechanisms of Cc Double Bond Reduction in Mammalian Fatty Acid Synthase Studied with Quantum Mechanics/Molecular Mechanics Calculations. *ACS Catal.* **2019**, *9*, 11404−11412.
(350) Pennacchio, A.; Giordano, A.; Esposito, L.; Langella, E.; Rossi, M.; Raia, C. A. Insight into the Stereospecificity of Short-Chain Thermus Thermophilus Alcohol Dehydrogenase Showing Pro-S Hydride Transfer and Prelog Enantioselectivity. *Protein Pept. Lett.* **2010**, *17*, 437−43.
(351) Libby, R. D.; Mehl, R. A. Characterization of Covalent Ene Adduct Intermediates in "Hydride Equivalent" Transfers in a Dihydropyridine Model for Nadh Reduction Reactions. *Bioorg. Chem.* **2012**, *40*, 57−66.
(352) Hamilton, G. A. Proton in Biological Redox Reactions. *Prog. Bioorg. Chem.* **1971**, *1*, 83−157.
(353) Donald, J. E.; Kulp, D. W.; DeGrado, W. F. Salt Bridges: Geometrically Specific, Designable Interactions. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 898−915.
(354) Grininger, M. Perspectives on the Evolution, Assembly and Conformational Dynamics of Fatty Acid Synthase Type I (Fas I) Systems. *Curr. Opin. Struct. Biol.* **2014**, *25*, 49−56.
(355) Lenfant, N.; Hotelier, T.; Bourne, Y.; Marchot, P.; Chatonnet, A. Proteins with an Alpha/Beta Hydrolase Fold: Relationships between Subfamilies in an Ever-Growing Superfamily. *Chem.-Biol. Interact.* **2013**, *203*, 266−8.
(356) Röttig, M.; Medema, M. H.; Blin, K.; Weber, T.; Rausch, C.; Kohlbacher, O. Nrpspredictor2—a Web Server for Predicting Nrps Adenylation Domain Specificity. *Nucleic Acids Res.* **2011**, *39*, W362−W367.
(357) Yu, X.; Liu, T.; Zhu, F.; Khosla, C. In Vitro Reconstitution and Steady-State Analysis of the Fatty Acid Synthase from Escherichia Coli. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 18643−18648.
(358) Yuzawa, S.; Kim, W.; Katz, L.; Keasling, J. D. Heterologous Production of Polyketides by Modular Type I Polyketide Synthases in Escherichia Coli. *Curr. Opin. Biotechnol.* **2012**, *23*, 727−735.
(359) Mattick, J. S.; Nickless, J.; Mizugaki, M.; Yang, C. Y.; Uchiyama, S.; Wakil, S. J. The Architecture of the Animal Fatty Acid Synthetase. Ii. Separation of the Core and Thioesterase Functions and Determination of the N-C Orientation of the Subunit. *J. Biol. Chem.* **1983**, *258*, 15300−4.
(360) Lin, C. Y.; Smith, S. Properties of the Thioesterase Component Obtained by Limited Trypsinization of the Fatty Acid Synthetase Multienzyme Complex. *J. Biol. Chem.* **1978**, *253*, 1954−62.
(361) Pazirandeh, M.; Chirala, S. S.; Huang, W. Y.; Wakil, S. J. Characterization of Recombinant Thioesterase and Acyl Carrier Protein Domains of Chicken Fatty Acid Synthase Expressed in Escherichia Coli. *J. Biol. Chem.* **1989**, *264*, 18195−201.
(362) Libertini, L. J.; Smith, S. Purification and Properties of a Thioesterase from Lactating Rat Mammary Gland Which Modifies the Product Specificity of Fatty Acid Synthetase. *J. Biol. Chem.* **1978**, *253*, 1393−401.
(363) Nolin, J. M.; Thompson, B. J.; Smith, S. Localization of Thioesterase Ii, the Chain-Length Regulatory Enzyme of Milk Fatty

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

Acid Synthesis, in Rat Mammary Gland Epithelial Cells. *J. Endocrinol.* **1982**, *94*, 251−6.
(364) Anderson, S. M.; Rudolph, M. C.; McManaman, J. L.; Neville, M. C. Key Stages in Mammary Gland Development. Secretory Activation in the Mammary Gland: It's Not Just About Milk Protein Synthesis! *Breast Cancer Res.* **2007**, *9*, 204.
(365) German, J. B. Dietary Lipids from an Evolutionary Perspective: Sources, Structures and Functions. *Matern. Child Nutr.* **2011**, *7*, 2−16.
(366) Ritchie, M. K.; Johnson, L. C.; Clodfelter, J. E.; Pemble, C. W.; Fulp, B. E.; Furdui, C. M.; Kridel, S. J.; Lowther, W. T.; et al. Crystal Structure and Substrate Specificity of Human Thioesterase 2: Insights into the Molecular Basis for the Modulation of Fatty Acid Synthase. *J. Biol. Chem.* **2016**, *291*, 3520−3530.
(367) Cantu, D. C.; Chen, Y.; Reilly, P. J. Thioesterases: A New Perspective Based on Their Primary and Tertiary Structures. *Protein Sci.* **2010**, *19*, 1281−1295.
(368) Ollis, D. L.; Cheah, E.; Cygler, M.; Dijkstra, B.; Frolow, F.; Franken, S. M.; Harel, M.; Remington, S. J.; Silman, I.; Schrag, J.; et al. The α/β Hydrolase Fold. *Protein Eng., Des. Sel.* **1992**, *5*, 197−211.
(369) Lenfant, N.; Hotelier, T.; Velluet, E.; Bourne, Y.; Marchot, P.; Chatonnet, A. Esther, the Database of the α/β-Hydrolase Fold Superfamily of Proteins: Tools to Explore Diversity of Functions. *Nucleic Acids Res.* **2012**, *41*, D423−D429.
(370) Korman, T. P.; Crawford, J. M.; Labonte, J. W.; Newman, A. G.; Wong, J.; Townsend, C. A.; Tsai, S.-C. Structure and Function of an Iterative Polyketide Synthase Thioesterase Domain Catalyzing Claisen Cyclization in Aflatoxin Biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 6246−6251.
(371) Gay, D. C.; Spear, P. J.; Keatinge-Clay, A. T. A Double-Hotdog with a New Trick: Structure and Mechanism of the Trans -Acyltransferase Polyketide Synthase Enoyl-Isomerase. *ACS Chem. Biol.* **2014**, *9*, 2374−2381.
(372) Liu, Y.; Zheng, T.; Bruner, S. D. Structural Basis for Phosphopantetheinyl Carrier Domain Interactions in the Terminal Module of Nonribosomal Peptide Synthetases. *Chem. Biol.* **2011**, *18*, 1482−1488.
(373) Huguenin-Dezot, N.; Alonzo, D. A.; Heberlig, G. W.; Mahesh, M.; Nguyen, D. P.; Dornan, M. H.; Boddy, C. N.; Schmeing, T. M.; Chin, J. W. Trapping Biosynthetic Acyl-Enzyme Intermediates with Encoded 2,3-Diaminopropionic Acid. *Nature* **2019**, *565*, 112−117.
(374) Bruner, S. D.; Weber, T.; Kohli, R. M.; Schwarzer, D.; Marahiel, M. A.; Walsh, C. T.; Stubbs, M. T. Structural Basis for the Cyclization of the Lipopeptide Antibiotic Surfactin by the Thioesterase Domain Srfte. *Structure* **2002**, *10*, 301−310.
(375) Samel, S. A.; Wagner, B.; Marahiel, M. A.; Essen, L. O. The Thioesterase Domain of the Fengycin Biosynthesis Cluster: A Structural Base for the Macrocyclization of a Non-Ribosomal Lipopeptide. *J. Mol. Biol.* **2006**, *359*, 876−889.
(376) Cheng, F.; Wang, Q.; Chen, M.; Quiocho, F. A.; Ma, J. Molecular Docking Study of the Interactions between the Thioesterase Domain of Human Fatty Acid Synthase and Its Ligands. *Proteins: Struct., Funct., Genet.* **2008**, *70*, 1228−1234.
(377) Giraldes, J. W.; Akey, D. L.; Kittendorf, J. D.; Sherman, D. H.; Smith, J. L.; Fecik, R. A. Structural and Mechanistic Insights into Polyketide Macrolactonization from Polyketide-Based Affinity Labels. *Nat. Chem. Biol.* **2006**, *2*, 531−536.
(378) Strauss, E. Coenzyme A Biosynthesis and Enzymology. *Comprehensive Natural Products II: Chemistry and Biology*; Elsevier, 2010; Vol. *7* pp 351410.
(379) Fako, V. E.; Zhang, J.-T.; Liu, J.-Y. Mechanism of Orlistat Hydrolysis by the Thioesterase of Human Fatty Acid Synthase. *ACS Catal.* **2014**, *4*, 3444−3453.
(380) Gulick, A. M.; Aldrich, C. C. Trapping Interactions between Catalytic Domains and Carrier Proteins of Modular Biosynthetic Enzymes with Chemical Probes. *Nat. Prod. Rep.* **2018**, *35*, 1156−1184.
(381) Denesyuk, A.; Dimitriou, P. S.; Johnson, M. S.; Nakayama, T.; Denessiouk, K. The Acid-Base-Nucleophile Catalytic Triad in Abh-Fold Enzymes Is Coordinated by a Set of Structural Elements. *PLoS One* **2020**, *15*, e0229376.
(382) Pazirandeh, M.; Chirala, S. S.; Wakil, S. J. Site-Directed Mutagenesis Studies on the Recombinant Thioesterase Domain of Chicken Fatty-Acid Synthase Expressed in Escherichia-Coli. *J. Biol. Chem.* **1991**, *266*, 20946−20952.
(383) Buller, A. R.; Townsend, C. A. Intrinsic Evolutionary Constraints on Protease Structure, Enzyme Acylation, and the Identity of the Catalytic Triad. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E653−E661.
(384) Koglin, A.; Löhr, F.; Bernhard, F.; Rogov, V. V.; Frueh, D. P.; Strieter, E. R.; Mofid, M. R.; Güntert, P.; Wagner, G.; Walsh, C. T.; et al. Structural Basis for the Selectivity of the External Thioesterase of the Surfactin Synthetase. *Nature* **2008**, *454*, 907−911.
(385) Claxton, H. B.; Akey, D. L.; Silver, M. K.; Admiraal, S. J.; Smith, J. L. Structure and Functional Analysis of Rifr, the Type Ii Thioesterase from the Rifamycin Biosynthetic Pathway. *J. Biol. Chem.* **2009**, *284*, 5021−9.
(386) Frueh, D. P.; Arthanari, H.; Koglin, A.; Vosburg, D. A.; Bennett, A. E.; Walsh, C. T.; Wagner, G. Dynamic Thiolation-Thioesterase Structure of a Non-Ribosomal Peptide Synthetase. *Nature* **2008**, *454*, 903−906.
(387) Tsai, S.-C.; Miercke, L. J.; Krucinski, J.; Gokhale, R.; Chen, J. C.; Foster, P. G.; Cane, D. E.; Khosla, C.; Stroud, R. M. Crystal Structure of the Macrocycle-Forming Thioesterase Domain of the Erythromycin Polyketide Synthase: Versatility from a Unique Substrate Channel. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 14808−13.
(388) Tsai, S.-C.; Lu, H.; Cane, D. E.; Khosla, C.; Stroud, R. M. Insights into Channel Architecture and Substrate Specificity from Crystal Structures of Two Macrocycle-Forming Thioesterases of Modular Polyketide Synthases. *Biochemistry* **2002**, *41*, 12598−606.
(389) Choi, Y. J.; Lee, S. Y. Microbial Production of Short-Chain Alkanes. *Nature* **2013**, *502*, 571−574.
(390) Feng, Y.; Zhang, Y.; Wang, Y.; Liu, J.; Liu, Y.; Cao, X.; Xue, S. Tuning of Acyl-Acp Thioesterase Activity Directed for Tailored Fatty Acid Synthesis. *Appl. Microbiol. Biotechnol.* **2018**, *102*, 3173−3182.
(391) Zheng, Y.-N.; Li, L.-L.; Liu, Q.; Yang, J.-M.; Wang, X.-W.; Liu, W.; Xu, X.; Liu, H.; Zhao, G.; Xian, M. Optimization of Fatty Alcohol Biosynthesis Pathway for Selectively Enhanced Production of C12/14 and C16/18 Fatty Alcohols in Engineered Escherichia Coli. *Microb. Cell Fact.* **2012**, *11*, 65.
(392) Xu, P.; Qiao, K. J.; Ahn, W. S.; Stephanopoulos, G. Engineering Yarrowia Lipolytica as a Platform for Synthesis of Drop-in Transportation Fuels and Oleochemicals. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 10848−10853.
(393) Agapakis, C. M.; Boyle, P. M.; Silver, P. A. Natural Strategies for the Spatial Optimization of Metabolism in Synthetic Biology. *Nat. Chem. Biol.* **2012**, *8*, 527−535.
(394) Dansen, T. B.; Wirtz, K. W. A.; Wanders, R. J. A.; Pap, E. H. W. Peroxisomes in Human Fibroblasts Have a Basic Ph. *Nat. Cell Biol.* **2000**, *2*, 51−53.
(395) Reubsaet, F. A.; Veerkamp, J. H.; Brückwilder, M. L.; Trijbels, J. M.; Monnens, L. A. Peroxisomal Oxidases and Catalase in Liver and Kidney Homogenates of Normal and Di(Ethylhexyl)Phthalate-Fed Rats. *Int. J. Biochem.* **1991**, *23*, 961−7.
(396) Antonenkov, V. D.; Grunau, S.; Ohlmeier, S.; Hiltunen, J. K. Peroxisomes Are Oxidative Organelles. *Antioxid. Redox Signaling* **2010**, *13*, 525−37.
(397) Sweetlove, L. J.; Fernie, A. R. The Role of Dynamic Enzyme Assemblies and Substrate Channelling in Metabolic Regulation. *Nat. Commun.* **2018**, *9*, 2136.
(398) Fulton, A. B. How Crowded Is the Cytoplasm? *Cell* **1982**, *30*, 345−347.
(399) Ovádi, J.; Saks, V. On the Origin of Intracellular Compartmentation and Organized Metabolic Systems. *Mol. Cell. Biochem.* **2004**, *256*, 5−12.
(400) Gaertner, F. H. Unique Catalytic Properties of Enzyme Clusters. *Trends Biochem. Sci.* **1978**, *3*, 63−65.
(401) Castellana, M.; Wilson, M. Z.; Xu, Y.; Joshi, P.; Cristea, I. M.; Rabinowitz, J. D.; Gitai, Z.; Wingreen, N. S. Enzyme Clustering

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

Accelerates Processing of Intermediates through Metabolic Channeling. *Nat. Biotechnol.* **2014**, *32*, 1011−1018.
(402) Wombacher, H. Molecular Compartmentation by Enzyme Cluster Formation. A View over Current Investigations. *Mol. Cell. Biochem.* **1983**, *56*, 155−164.
(403) Bukhari, H. S. T.; Jakob, R. P.; Maier, T. Evolutionary Origins of the Multienzyme Architecture of Giant Fungal Fatty Acid Synthase. *Structure* **2014**, *22*, 1775−1785.
(404) Lieber, D. J.; Catlett, J.; Madayiputhiya, N.; Nandakumar, R.; Lopez, M. M.; Metcalf, W. W.; Buan, N. R. A Multienzyme Complex Channels Substrates and Electrons through Acetyl-Coa and Methane Biosynthesis Pathways in Methanosarcina. *PLoS One* **2014**, *9*, No. e107563.
(405) Abernathy, M. H.; He, L.; Tang, Y. J. Channeling in Native Microbial Pathways: Implications and Challenges for Metabolic Engineering. *Biotechnol. Adv.* **2017**, *35*, 805−814.
(406) Pettersson, G. Evolutionary Optimization of the Catalytic Efficiency of Enzymes. *Eur. J. Biochem.* **1992**, *206*, 289−295.
(407) Srere, P. A. In *Energy Metabolism and the Regulation of Metabolic Processes in Mitochondria*; Mehlman, M. A., Hanson, R. W., Eds.; Academic Press, 1972. DOI: 10.1016/B978-0-12-487850-1.50011-7.
(408) Fischer, M.; Grininger, M. Strategies in Megasynthase Engineering - Fatty Acid Synthases (Fas) as Model Proteins. *Beilstein J. Org. Chem.* **2017**, *13*, 1204−1211.
(409) Ishihama, Y.; Schmidt, T.; Rappsilber, J.; Mann, M.; Hartl, F. U.; Kerner, M. J.; Frishman, D. Protein Abundance Profiling of the Escherichia Coli Cytosol. *BMC Genomics* **2008**, *9*, 102.
(410) Finking, R.; Marahiel, M. A. Biosynthesis of Nonribosomal Peptides. *Annu. Rev. Microbiol.* **2004**, *58*, 453−488.
(411) Weissman, K. J. Polyketide Biosynthesis: Understanding and Exploiting Modularity. *Philos. Trans. R. Soc., A* **2004**, *362*, 2671−2690.
(412) Wong, H. C.; Liu, G.; Zhang, Y.-M.; Rock, C. O.; Zheng, J. The Solution Structure of Acyl Carrier Protein from Mycobacterium Tuberculosis. *J. Biol. Chem.* **2002**, *277*, 15874−15880.
(413) Sharma, A. K.; Sharma, S. K.; Surolia, A.; Surolia, N.; Sarma, S. P. Solution Structures of Conformationally Equilibrium Forms of Holo-Acyl Carrier Protein (Pfacp) from Plasmodium Falciparum Provides Insight into the Mechanism of Activation of Acps. *Biochemistry* **2006**, *45*, 6904−6916.
(414) Kim, Y.; Kovrigin, E. L.; Eletr, Z. Nmr Studies of Escherichia Coli Acyl Carrier Protein: Dynamic and Structural Differences of the Apo- and Holo-Forms. *Biochem. Biophys. Res. Commun.* **2006**, *341*, 776−783.
(415) Bloch, K.; Vance, D. Control Mechanisms in the Synthesis of Saturated Fatty Acids. *Annu. Rev. Biochem.* **1977**, *46*, 263−298.
(416) Heath, R. J.; Jackowski, S.; Rock, C. O. Fatty acid and phospholipid metabolism in prokaryotes. *New Comprehensive Biochemistry*, Vol. 36; Elsevier: Amsterdam, 2002. DOI: 10.1016/S0167-7306(02)36005-8.
(417) Butland, G.; Peregrín-Alvarez, J. M.; Li, J.; Yang, W.; Yang, X.; Canadien, V.; Starostine, A.; Richards, D.; Beattie, B.; Krogan, N.; et al. Interaction Network Containing Conserved and Essential Protein Complexes in *Escherichia Coli*. *Nature* **2005**, *433*, 531−537.
(418) Rock, C. O.; Cronan, J. E. *Methods in Enzymology*; Academic Press, 1981; Vol. *71*.
(419) Lu, P.; Vogel, C.; Wang, R.; Yao, X.; Marcotte, E. M. Absolute Protein Expression Profiling Estimates the Relative Contributions of Transcriptional and Translational Regulation. *Nat. Biotechnol.* **2007**, *25*, 117−124.
(420) Fichtlscherer, F.; Wellein, C.; Mittag, M.; Schweizer, E. A Novel Function of Yeast Fatty Acid Synthase. Subunit A Is Capable of Self-Pantetheinylation. *Eur. J. Biochem.* **2000**, *267*, 2666−2671.
(421) Jayakumar, A.; Tai, M. H.; Huang, W. Y.; Al-Feel, W.; Hsu, M.; Abu-Elheiga, L.; Chirala, S. S.; Wakil, S. J. Human Fatty Acid Synthase: Properties and Molecular Cloning. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92*, 8695−9.
(422) Stuible, H. P.; Meurer, G.; Schweizer, E. Heterologous Expression and Biochemical Characterization of Two Functionally Different Type I Fatty Acid Synthases from Brevibacterium Ammoniagenes. *Eur. J. Biochem.* **1997**, *247*, 268−273.
(423) Nivina, A.; Yuet, K. P.; Hsu, J.; Khosla, C. Evolution and Diversity of Assembly-Line Polyketide Synthases. *Chem. Rev.* **2019**, *119*, 12524−12547.
(424) Chen, A. C.; Re, R. N.; Burkart, M. D. Type Ii Fatty Acid and Polyketide Synthases: Deciphering Protein-Protein and Protein-Substrate Interactions. *Nat. Prod. Rep.* **2018**, *35*, 1029−1045.
(425) Dodge, G. J.; Maloney, F. P.; Smith, J. L. Protein-Protein Interactions in "Cis-at" Polyketide Synthases. *Nat. Prod. Rep.* **2018**, *35*, 1082−1096.
(426) Du, D.; Katsuyama, Y.; Shin-Ya, K.; Ohnishi, Y. Reconstitution of a Type Ii Polyketide Synthase That Catalyzes Polyene Formation. *Angew. Chem., Int. Ed.* **2018**, *57*, 1954−1957.
(427) Klaus, M.; Grininger, M. Engineering Strategies for Rational Polyketide Synthase Design. *Nat. Prod. Rep.* **2018**, *35*, 1070−1081.
(428) Kosol, S.; Jenner, M.; Lewandowski, J. R.; Challis, G. L. Protein-Protein Interactions in Trans-at Polyketide Synthases. *Nat. Prod. Rep.* **2018**, *35*, 1097−1109.
(429) Poust, S.; Hagen, A.; Katz, L.; Keasling, J. D. Narrowing the Gap between the Promise and Reality of Polyketide Synthases as a Synthetic Biology Platform. *Curr. Opin. Biotechnol.* **2014**, *30*, 32−39.
(430) Weissman, K. J. Peering into the Black Box of Fungal Polyketide Biosynthesis. *ChemBioChem* **2010**, *11*, 485−488.
(431) Das, A.; Khosla, C. Biosynthesis of Aromatic Polyketides in Bacteria. *Acc. Chem. Res.* **2009**, *42*, 631−639.
(432) Cropp, T. A.; Kim, B. S.; Beck, B. J.; Yoon, Y. J.; Sherman, D. H.; Reynolds, K. A. Recent Developments in the Production of Novel Polyketides by Combinatorial Biosynthesis. *Biotechnol. Genet. Eng. Rev.* **2002**, *19*, 159−172.
(433) Staunton, J.; Weissman, K. J. Polyketide Biosynthesis: A Millennium Review. *Nat. Prod. Rep.* **2001**, *18*, 380−416.
(434) Cane, D. E.; Walsh, C. T.; Khosla, C. Biochemistry - Harnessing the Biosynthetic Code: Combinations, Permutations, and Mutations. *Science* **1998**, *282*, 63−68.
(435) Moore, B. S.; Hertweck, C. Biosynthesis and Attachment of Novel Bacterial Polyketide Synthase Starter Units. *Nat. Prod. Rep.* **2002**, *19*, 70−99.
(436) Zhou, H.; Zhan, J. X.; Watanabe, K.; Xie, X. K.; Tang, Y. A Polyketide Macrolactone Synthase from the Filamentous Fungus Gibberella Zeae. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 6249−6254.
(437) Straight, P. D.; Fischbach, M. A.; Walsh, C. T.; Rudner, D. Z.; Kolter, R. A Singular Enzymatic Megacomplex from Bacillus Subtilis. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 305−310.
(438) Oyola-Robles, D.; Gay, D. C.; Trujillo, U.; Sanchez-Pares, J. M.; Bermudez, M. L.; Rivera-Diaz, M.; Carballeira, N. M.; Baerga-Ortiz, A. Identification of Novel Protein Domains Required for the Expression of an Active Dehydratase Fragment from a Polyunsaturated Fatty Acid Synthase. *Protein Sci.* **2013**, *22*, 954−963.
(439) Kohli, G. S.; John, U.; Van Dolah, F. M.; Murray, S. A. Evolutionary Distinctiveness of Fatty Acid and Polyketide Synthesis in Eukaryotes. *ISME J.* **2016**, *10*, 1877−1890.
(440) Pawlowiez, R.; Morey, J. S.; Darius, H. T.; Chinain, M.; Van Dolah, F. M. Transcriptome Sequencing Reveals Single Domain Type I-Like Polyketide Synthases in the Toxic Dinoflagellate Gambierdiscus Polynesiensis. *Harmful Algae* **2014**, *36*, 29−37.
(441) Kohli, G. S.; John, U.; Figueroa, R. I.; Rhodes, L. L.; Harwood, D. T.; Groth, M.; Bolch, C. J. S.; Murray, S. A. Polyketide Synthesis Genes Associated with Toxin Production in Two Species of Gambierdiscus (Dinophyceae). *BMC Genomics* **2015**, *16*, 410.
(442) Yu, D. Y.; Xu, F. C.; Zeng, J.; Zhan, J. X. Type Iii Polyketide Synthases in Natural Product Biosynthesis. *IUBMB Life* **2012**, *64*, 285−295.
(443) Muller, R. Don't Classify Polyketide Synthases. *Chem. Biol.* **2004**, *11*, 4−6.
(444) Minotti, G.; Menna, P.; Salvatorelli, E.; Cairo, G.; Gianni, L. Anthracyclines: Molecular Advances and Pharmacologic Developments in Antitumor Activity and Cardiotoxicity. *Pharmacol. Rev.* **2004**, *56*, 185−229.

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

(445) Griffin, M. O.; Fricovsky, E.; Ceballos, G.; Villarreal, F. Tetracyclines: A Pleitropic Family of Compounds with Promising Therapeutic Properties. Review of the Literature. *Am. J. Physiol-Cell Ph.* **2010**, *299*, C539−C548.
(446) Mcdaniel, R.; Ebertkhosla, S.; Hopwood, D. A.; Khosla, C. Engineered Biosynthesis of Novel Polyketides - Act(Vii) and Act(Iv) Genes Encode Aromatase and Cyclase Enzymes, Respectively. *J. Am. Chem. Soc.* **1994**, *116*, 10855−10859.
(447) Taguchi, T.; Awakawa, T.; Nishihara, Y.; Kawamura, M.; Ohnishi, Y.; Ichinose, K. Bifunctionality of Activ as a Cyclase-Thioesterase Revealed by in Vitro Reconstitution of Actinorhodin Biosynthesis in Streptomyces Coelicolor A3(2). *ChemBioChem* **2017**, *18*, 316−323.
(448) Shimizu, Y.; Ogata, H.; Goto, S. Type Iii Polyketide Synthases: Functional Classification and Phylogenomics. *ChemBioChem* **2017**, *18*, 50−65.
(449) Shimizu, Y.; Ogata, H.; Goto, S. Type Iii Polyketide Synthases: Functional Classification and Phylogenomics (Vol 18, Pg 50, 2017). *ChemBioChem* **2017**, *18*, 1048−1049.
(450) Liu, B.; Raeth, T.; Beuerle, T.; Beerhues, L. Biphenyl Synthase, a Novel Type Iii Polyketide Synthase. *Planta* **2007**, *225*, 1495−1503.
(451) Ferrer, J. L.; Jez, J. M.; Bowman, M. E.; Dixon, R. A.; Noel, J. P. Structure of Chalcone Synthase and the Molecular Basis of Plant Polyketide Biosynthesis. *Nat. Struct. Biol.* **1999**, *6*, 775−784.
(452) Jez, J. M.; Bowman, M. E.; Noel, J. P. Expanding the Biosynthetic Repertoire of Plant Type Iii Polyketide Synthases by Altering Starter Molecule Specificity. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 5319−5324.
(453) Song, L. J.; Barona-Gomez, F.; Corre, C.; Xiang, L. K.; Udwary, D. W.; Austin, M. B.; Noel, J. P.; Moore, B. S.; Challis, G. L. Type Iii Polyketide Synthase Beta-Ketoacyl-Acp Starter Unit and Ethylmalonyl-Coa Extender Unit Selectivity Discovered by Streptomyces Coelicolor Genome Mining. *J. Am. Chem. Soc.* **2006**, *128*, 14754−14755.
(454) Funabashi, M.; Funa, N.; Horinouchi, S. Phenolic Lipids Synthesized by Type Iii Polyketide Synthase Confer Penicillin Resistance on Streptomyces Griseus. *J. Biol. Chem.* **2008**, *283*, 25104−25104.
(455) Austin, M. B.; Noel, A. J. P. The Chalcone Synthase Superfamily of Type Iii Polyketide Synthases. *Nat. Prod. Rep.* **2003**, *20*, 79−110.
(456) Newman, A. G.; Vagstad, A. L.; Storm, P. A.; Townsend, C. A. Systematic Domain Swaps of Iterative, Nonreducing Polyketide Synthases Provide a Mechanistic Understanding and Rationale for Catalytic Reprogramming. *J. Am. Chem. Soc.* **2014**, *136*, 7348−7362.
(457) Fisch, K. M.; Bakeer, W.; Yakasai, A. A.; Song, Z. S.; Pedrick, J.; Wasil, Z.; Bailey, A. M.; Lazarus, C. M.; Simpson, T. J.; Cox, R. J. Rational Domain Swaps Decipher Programming in Fungal Highly Reducing Polyketide Synthases and Resurrect an Extinct Metabolite. *J. Am. Chem. Soc.* **2011**, *133*, 16635−16641.
(458) Barajas, J. F.; Blake-Hedges, J. M.; Bailey, C. B.; Curran, S.; Keasling, J. D. Engineered Polyketides: Synergy between Protein and Host Level Engineering. *Synth. Syst. Biotechno.* **2017**, *2*, 147−166.
(459) Mcdaniel, R.; Ebertkhosla, S.; Hopwood, D. A.; Khosla, C. Rational Design of Aromatic Polyketide Natural-Products by Recombinant Assembly of Enzymatic Subunits. *Nature* **1995**, *375*, 549−554.
(460) Menzella, H. G.; Reid, R.; Carney, J. R.; Chandran, S. S.; Reisinger, S. J.; Patel, K. G.; Hopwood, D. A.; Santi, D. V. Combinatorial Polyketide Biosynthesis by De Novo Design and Rearrangement of Modular Polyketide Synthase Genes. *Nat. Biotechnol.* **2005**, *23*, 1171−1176.
(461) Miyazawa, T.; Hirsch, M.; Zhang, Z. C.; Keatinge-Clay, A. T. An in Vitro Platform for Engineering and Harnessing Modular Polyketide Synthases. *Nat. Commun.* **2020**, *11*, 80.
(462) Morita, H.; Wong, C. P.; Abe, I. How Structural Subtleties Lead to Molecular Diversity for the Type Iii Polyketide Synthases. *J. Biol. Chem.* **2019**, *294*, 15121−15136.
(463) Knothe, G. Biodiesel and Renewable Diesel: A Comparison. *Prog. Energy Combust. Sci.* **2010**, *36*, 364−373.
(464) Hu, Y. T.; Zhu, Z. W.; Nielsen, J.; Siewers, V. Engineering Saccharomyces Cerevisiae Cells for Production of Fatty Acid-Derived Biofuels and Chemicals. *Open Biol.* **2019**, *9*, 190049.
(465) Rigouin, C.; Croux, C.; Borsenberger, V.; Ben Khaled, M.; Chardot, T.; Marty, A.; Bordes, F. Increasing Medium Chain Fatty Acids Production in Yarrowia Lipolytica by Metabolic Engineering. *Microb. Cell Fact.* **2018**, *17*, 142.
(466) Xue, P.; Si, T.; Mishra, S.; Zhang, L. Z. X.; Choe, K.; Sweedler, J. V.; Zhao, H. M. A Mass Spectrometry-Based High-Throughput Screening Method for Engineering Fatty Acid Synthases with Improved Production of Medium-Chain Fatty Acids. *Biotechnol. Bioeng.* **2020**, *117*, 2131−2138.
(467) Marella, E. R.; Holkenbrink, C.; Siewers, V.; Borodina, I. Engineering Microbial Fatty Acid Metabolism for Biofuels and Biochemicals. *Curr. Opin. Biotechnol.* **2018**, *50*, 39−46.
(468) Runguphan, W.; Keasling, J. D. Metabolic Engineering of Saccharomyces Cerevisiae for Production of Fatty Acid-Derived Biofuels and Chemicals. *Metab. Eng.* **2014**, *21*, 103−113.
(469) Cronan, J. E.; Weisberg, L. J.; Allen, R. G. Regulation of Membrane Lipid-Synthesis in *Escherichia-Coli* - Accumulation of Free Fatty-Acids of Abnormal Length During Inhibition of Phospholipid Synthesis. *J. Biol. Chem.* **1975**, *250*, 5835−5840.
(470) Kalscheuer, R.; Stoveken, T.; Luftmann, H.; Malkus, U.; Reichelt, R.; Steinbuchel, A. Neutral Lipid Biosynthesis in Engineered Escherichia Coli: Jojoba Oil-Like Wax Esters and Fatty Acid Butyl Esters. *Appl. Environ. Microbiol.* **2006**, *72*, 1373−1379.
(471) Steen, E. J.; Kang, Y. S.; Bokinsky, G.; Hu, Z. H.; Schirmer, A.; McClure, A.; del Cardayre, S. B.; Keasling, J. D. Microbial Production of Fatty-Acid-Derived Fuels and Chemicals from Plant Biomass. *Nature* **2010**, *463*, 559−U182.
(472) Keasling, J. D.; Kang, Y.; Steen, E. J.; Bokinsky, G. Product of Fatty Acid Esters from Biomass Polymers. WO 2010/127318 A2, 4 Nov 2010.
(473) Ruhdal Jensen, P. Production of Biodiesel by Yeast from Lignocellulose and Glycerol. WO 2011/157848 Al, 22 Dec 2011.
(474) Handke, P.; Lynch, S. A.; Gill, R. T. Application and Engineering of Fatty Acid Biosynthesis in Escherichia Coli for Advanced Fuels and Chemicals. *Metab. Eng.* **2011**, *13*, 28−37.
(475) Xu, P.; Gu, Q.; Wang, W. Y.; Wong, L.; Bower, A. G. W.; Collins, C. H.; Koffas, M. A. G. Modular Optimization of Multi-Gene Pathways for Fatty Acids Production in E. Coli. *Nat. Commun.* **2013**, *4*, 1409.
(476) Lennen, R. M.; Braden, D. J.; West, R. M.; Dumesic, J. A.; Pfleger, B. F. A Process for Microbial Hydrocarbon Synthesis: Overproduction of Fatty Acids in Escherichia Coli and Catalytic Conversion to Alkanes. *Biotechnol. Bioeng.* **2010**, *106*, 193−202.
(477) Łos, M.; Czyz, A.; Sell, E.; Wegrzyn, A.; Neubauer, P.; Wegrzyn, G. Bacteriophage Contamination: Is There a Simple Method to Reduce Its Deleterious Effects in Laboratory Cultures and Biotechnological Factories? *J. Appl. Genet.* **2004**, *45*, 111−120.
(478) Zhu, Z. W.; Hu, Y. T.; Teixeira, P. G.; Pereira, R.; Chen, Y.; Siewers, V.; Nielsen, J. Multidimensional Engineering of Saccharomyces Cerevisiae for Efficient Synthesis of Medium-Chain Fatty Acids. *Nat. Catal.* **2020**, *3*, 64−74.
(479) Michinaka, Y.; Shimauchi, T.; Aki, T.; Nakajima, T.; Kawamoto, S.; Shigeta, S.; Suzuki, O.; Ono, K. Extracellular Secretion of Free Fatty Acids by Disruption of a Fatty Acyl-Coa Synthetase Gene in Saccharomyces Cerevisiae. *J. Biosci. Bioeng.* **2003**, *95*, 435−440.
(480) Henritzi, S.; Fischer, M.; Grininger, M.; Oreb, M.; Boles, E. An Engineered Fatty Acid Synthase Combined with a Carboxylic Acid Reductase Enables De Novo Production of 1-Octanol in Saccharomyces Cerevisiae. *Biotechnol. Biofuels* **2018**, *11*, 150.
(481) Coursolle, D.; Lian, J. Z.; Shanklin, J.; Zhao, H. M. Production of Long Chain Alcohols and Alkanes Upon Coexpression of an Acyl-Acp Reductase and Aldehyde-Deformylating Oxygenase with a Bacterial Type-I Fatty Acid Synthase in *E. Coli*. *Mol. BioSyst.* **2015**, *11*, 2464−2472.
(482) Schirmer, A.; Rude, M. A.; Li, X. Z.; Popova, E.; del Cardayre, S. B. Microbial Biosynthesis of Alkanes. *Science* **2010**, *329*, 559−562.

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

(483) Haushalter, R. W.; Groff, D.; Deutsch, S.; The, L.; Chavkin, T. A.; Brunner, S. F.; Katz, L.; Keasling, J. D. Development of an Orthogonal Fatty Acid Biosynthesis System in E. Coli for Oleochemical Production. *Metab. Eng.* **2015**, *30*, 1−6.
(484) Eriksen, D. T.; HamediRad, M.; Yuan, Y. B.; Zhao, H. M. Orthogonal Fatty Acid Biosynthetic Pathway Improves Fatty Acid Ethyl Ester Production in Saccharomyces Cerevisiae. *ACS Synth. Biol.* **2015**, *4*, 808−814.
(485) Shin, G. H.; Veen, M.; Stahl, U.; Lang, C. Overexpression of Genes of the Fatty Acid Biosynthetic Pathway Leads to Accumulation of Sterols in Saccharomyces Cerevisiae. *Yeast* **2012**, *29*, 371−383.
(486) Milke, L.; Marienhagen, J. Engineering Intracellular Malonyl-Coa Availability in Microbial Hosts and Its Impact on Polyketide and Fatty Acid Synthesis. *Appl. Microbiol. Biotechnol.* **2020**, *104*, 6057−6065.
(487) Greenfield, D. L.; Trimbur, D. E.; Schirmer, A. W.; Chang, C.; Behrouzian, B.; Winger, J. Improved Acetyl-Coa Carboxylase Variants. WO 2015/038970 A2, 19 Mar 2015.
(488) Lo, Y. C.; Lin, S. C.; Shaw, J. F.; Liaw, Y. C. Substrate Specificities of Escherichia Coli Thioesterase I/Protease I/Lysophospholipase L-1 Are Governed by Its Switch Loop Movement. *Biochemistry* **2005**, *44*, 1971−1979.
(489) Ramesha, C.; Joshi, A. K.; Subramaniam, R. Compositions and Methods for Production of Biofuels. WO 2009/103065 A2, 20 Aug 2009.
(490) Gary, J. H.; Handwerk, G. E. *Petroleum Refining: Technology and Economics*, 4th ed.; Taylor & Francis, 2001.
(491) Jeon, E.; Lee, S.; Won, J. I.; Han, S. O.; Kim, J.; Lee, J. Development of Escherichia Coli Mg1655 Strains to Produce Long Chain Fatty Acids by Engineering Fatty Acid Synthesis (Fas) Metabolism. *Enzyme Microb. Technol.* **2011**, *49*, 44−51.
(492) Lee, S.; Jung, Y.; Lee, S.; Lee, J. Correlations between Fas Elongation Cycle Genes Expression and Fatty Acid Production for Improvement of Long-Chain Fatty Acids in Escherichia Coli. *Appl. Biochem. Biotechnol.* **2013**, *169*, 1606−1619.
(493) Jiang, W.; Jiang, Y. F.; Bentley, G. J.; Liu, D.; Xiao, Y.; Zhang, F. Z. Enhanced Production of Branched-Chain Fatty Acids by Replacing -Ketoacyl-(Acyl-Carrier-Protein) Synthase Iii (Fabh). *Biotechnol. Bioeng.* **2015**, *112*, 1613−1622.
(494) Knothe, G.; Dunn, R. O. A Comprehensive Evaluation of the Melting Points of Fatty Acids and Esters Determined by Differential Scanning Calorimetry. *J. Am. Oil Chem. Soc.* **2009**, *86*, 843−856.
(495) Zhang, F.; Jiang, W.; Bentley, G. Host Cells and Methods for Producing Fatty Acid-Derivatives with High Branched-Chain Percentage. WO 2017/106747 Al, 22 Jun 2017.
(496) Ruppe, A.; Fox, J. M. Analysis of Interdependent Kinetic Controls of Fatty Acid Synthases. *ACS Catal.* **2018**, *8*, 11722−11734.
(497) Wernig, F.; Born, S.; Boles, E.; Grininger, M.; Oreb, M. Fusing Alpha and Beta Subunits of the Fungal Fatty Acid Synthase Leads to Improved Production of Fatty Acids. *Sci. Rep.* **2020**, *10*, 9780.
(498) Runguphan, W.; Keasling, J. D. A Yeast Cell Modified to Overproduce Fatty Acid and Fatty Acid-Derived Compounds. WO 2015/013674 A2, 29 Jan 2015.
(499) Volker, A. R.; Gogerty, D. S.; Bartholomay, C.; Hennen-Bierwagen, T.; Zhu, H. L.; Bobik, T. A. Fermentative Production of Short-Chain Fatty Acids in Escherichia Coli. *Microbiology* **2014**, *160*, 1513−1522.
(500) Rodriguez, G. M.; Tashiro, Y.; Atsumi, S. Expanding Ester Biosynthesis in Escherichia Coli. *Nat. Chem. Biol.* **2014**, *10*, 259−265.
(501) Rutter, C. D.; Rao, C. V. Production of 1-Decanol by Metabolically Engineered Yarrowia Lipolytica. *Metab. Eng.* **2016**, *38*, 139−147.
(502) Kremer, F.; Blank, L. M.; Jones, P. R.; Akhtar, M. K. A Comparison of the Microbial Production and Combustion Characteristics of Three Alcohol Biofuels: Ethanol, 1-Butanol, and 1-Octanol. *Front. Bioeng. Biotechnol.* **2015**, *3*, 112.
(503) Kumar, B. R.; Saravanan, S.; Rana, D.; Nagendran, A. A Comparative Analysis on Combustion and Emissions of Some Next Generation Higher-Alcohol/Diesel Blends in a Direct-Injection Diesel Engine. *Energy Convers. Manage.* **2016**, *119*, 246−256.
(504) Borrull, A.; Lopez-Martinez, G.; Poblet, M.; Cordero-Otero, R.; Rozes, N. New Insights into the Toxicity Mechanism of Octanoic and Decanoic Acids on *Saccharomyces Cerevisiae*. *Yeast* **2015**, *32*, 451−460.
(505) Akita, H.; Matsukura, H.; Oishi, T. Lipase Catalyzed Enantioselective Hydrolysis of 2-Methyl 3-Acetoxy Esters. *Tetrahedron Lett.* **1986**, *27*, 5241−5244.
(506) Itoh, T.; Kuroda, K.; Tomosada, M.; Takagi, Y. Design of Alpha-Alkyl Beta-Hydroxy Esters Suitable for Providing Optical Resolution by Lipase Hydrolysis. *J. Org. Chem.* **1991**, *56*, 797−804.
(507) Rodriguez, S.; Kayser, M. M.; Stewart, J. D. Highly Stereoselective Reagents for Beta-Keto Ester Reductions by Genetic Engineering of Baker's Yeast. *J. Am. Chem. Soc.* **2001**, *123*, 1547−1555.
(508) Engelking, H.; Pfaller, R.; Wich, G.; Weuster-Botz, D. Reaction Engineering Studies on (-Ketoester Reductions with Whole Cells of Recombinant *Saccharomyces Cerevisiae*. *Enzyme Microb. Technol.* **2006**, *38*, 536−544.
(509) Davis, M. S.; Cronan, J. E. Inhibition of *Escherichia Coli* Acetyl Coenzyme A Carboxylase by Acyl-Acyl Carrier Protein. *J. Bacteriol.* **2001**, *183*, 1499−1503.
(510) de Jong, B. W.; Shi, S.; Siewers, V.; Nielsen, J. Improved Production of Fatty Acid Ethyl Esters in *Saccharomyces Cerevisiae* through up-Regulation of the Ethanol Degradation Pathway and Expression of the Heterologous Phosphoketolase Pathway. *Microb. Cell Fact.* **2014**, *13*, 39.
(511) Thompson, R. A.; Trinh, C. T. Enhancing Fatty Acid Ethyl Ester Production in Saccharomyces Cerevisiae through Metabolic Engineering and Medium Optimization. *Biotechnol. Bioeng.* **2014**, *111*, 2200−2208.
(512) Shi, S. B.; Chen, Y.; Siewers, V.; Nielsen, J. Improving Production of Malonyl Coenzyme A-Derived Metabolites by Abolishing Snf1-Dependent Regulation of Acc1. *mBio* **2014**, *5*, No. e01130-14.
(513) Valle-Rodriguez, J. O.; Shi, S. B.; Siewers, V.; Nielsen, J. Metabolic Engineering of Saccharomyces Cerevisiae for Production of Fatty Acid Ethyl Esters, an Advanced Biofuel, by Eliminating Non-Essential Fatty Acid Utilization Pathways. *Appl. Energy* **2014**, *115*, 226−232.
(514) David, F.; Siewers, V.; Krivoruchko, A.; Wenning, L. Fungal Cells and Methods for Production of Very Long Chain Fatty Acid Derived Products. WO 2016/159869 A1, 6 Oct 2016.
(515) Simpson, D.; da Costa, B.; Rude, M.; Trinh, N.; Popova, E.; Kiteswaran, S.; Helman, N. Acp-Mediated Production of Fatty Acid Derivatives. WO 2014/093505 A2, 19 Jun 2014.
(516) Aprahamian, S. A.; Arslanian, M. J.; Wakil, S. J. Comparative Studies on the Kinetic-Parameters and Product Analyses of Chicken and Rat-Liver and Yeast Fatty-Acid Synthetase. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **1982**, *71*, 577−582.
(517) Schweizer, E.; Hofmann, J. Microbial Type I Fatty Acid Synthases (Fas): Major Players in a Network of Cellular Fas Systems. *Microbiol. Mol. Biol. Rev.* **2004**, *68*, 501−517.
(518) Fernandez-Moya, R.; Leber, C.; Cardenas, J.; Da Silva, N. A. Functional Replacement of the Saccharomyces Cerevisiae Fatty Acid Synthase with a Bacterial Type Ii System Allows Flexible Product Profiles. *Biotechnol. Bioeng.* **2015**, *112*, 2618−2623.
(519) Lian, J. Z.; Zhao, H. M. Functional Reconstitution of a Pyruvate Dehydrogenase in the Cytosol of Saccharomyces Cerevisiae through Lipoylation Machinery Engineering. *ACS Synth. Biol.* **2016**, *5*, 689−697.
(520) Zhu, Z. W.; Zhou, Y. J. J.; Krivoruchko, A.; Grininger, M.; Zhao, Z. B. K.; Nielsen, J. Expanding the Product Portfolio of Fungal Type I Fatty Acid Synthases. *Nat. Chem. Biol.* **2017**, *13*, 360−362.
(521) Zheng, Y.; Li, L.; Liu, Q.; Yang, J.; Cao, Y.; Jiang, X.; Zhao, G.; Xian, M. Boosting the Free Fatty Acid Synthesis of Escherichia Coli by Expression of a Cytosolic Acinetobacter Baylyi Thioesterase. *Biotechnol. Biofuels* **2012**, *5*, 76.

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

(522) Shin, K. S.; Kim, S.; Lee, S. K. Improvement of Free Fatty Acid Production Using a Mutant Acyl-Coa Thioesterase I with High Specific Activity in Escherichia Coli. *Biotechnol. Biofuels* **2016**, *9*, 208.
(523) Gajewski, J.; Pavlovic, R.; Fischer, M.; Boles, E.; Grininger, M. Engineering Fungal De Novo Fatty Acid Synthesis for Short Chain Fatty Acid Production. *Nat. Commun.* **2017**, *8*, 14650.
(524) Grininger, M.; Gajewski, J.; Boles, E. Microbiological Production of Short Fatty Acids and Uses Thereof. WO 2017/001333 Al, 5 Jan 2017.
(525) Rigouin, C.; Gueroult, M.; Croux, C.; Dubois, G.; Borsenberger, V.; Barbe, S.; Marty, A.; Daboussi, F.; Andre, I.; Bordes, F. Production of Medium Chain Fatty Acids by Yarrowia Lipolytica: Combining Molecular Design and Talen to Engineer the Fatty Acid Synthase. *ACS Synth. Biol.* **2017**, *6*, 1870−1879.
(526) Leber, C.; Choi, J. W.; Polson, B.; Da Silva, N. A. Disrupted Short Chain Specific Beta-Oxidation and Improved Synthase Expression Increase Synthesis of Short Chain Fatty Acids in Saccharomyces Cerevisiae. *Biotechnol. Bioeng.* **2016**, *113*, 895−900.
(527) Williams, M. D.; Rahn, J. A.; Sherman, D. H. Production of a Polyhydroxyalkanoate Biopolymer in Insect Cells with a Modified Eucaryotic Fatty Acid Synthase. *Appl. Environ. Microbiol.* **1996**, *62*, 2540−2546.
(528) Abidi, S. H. I. Microbial Engineering for the Production of Cannabinoids and Cannabinoid Precursors. WO 2017/139496 A1, 17 Aug 2017.
(529) Gajewski, J.; Buelens, F.; Serdjukow, S.; Janssen, M.; Cortina, N.; Grubmuller, H.; Grininger, M. Engineering Fatty Acid Synthases for Directed Polyketide Production. *Nat. Chem. Biol.* **2017**, *13*, 363−365.
(530) Wang, F.; Wang, Y. J.; Ji, J. J.; Zhou, Z.; Yu, J. K.; Zhu, H.; Su, Z. G.; Zhang, L. X.; Zheng, J. T. Structural and Functional Analysis of the Loading Acyltransferase from Avermectin Modular Polyketide Synthase. *ACS Chem. Biol.* **2015**, *10*, 1017−1025.
(531) Wesener, S. R.; Potharla, V. Y.; Cheng, Y. Q. Reconstitution of the Fk228 Biosynthetic Pathway Reveals Cross Talk between Modular Polyketide Synthases and Fatty Acid Synthase. *Appl. Environ. Microbiol.* **2011**, *77*, 1501−1507.
(532) Revill, W. P.; Bibb, M. J.; Hopwood, D. A. Purification of a Malonyltransferase from Streptomyces-Coelicolor A3(2) and Analysis of Its Genetic Determinant. *J. Bacteriol.* **1995**, *177*, 3946−3952.
(533) Summers, R. G.; Ali, A.; Shen, B.; Wessel, W. A.; Hutchinson, C. R. Malonyl-Coenzyme A-Acyl Carrier Protein Acyltransferase of Streptomyces-Glaucescens - a Possible Link between Fatty-Acid and Polyketide Biosynthesis. *Biochemistry* **1995**, *34*, 9389−9402.
(534) Garg, S.; Rizhsky, L.; Jin, H. N.; Yu, X. C.; Jing, F. Y.; Yandeau-Nelson, M. D.; Nikolau, B. J. Microbial Production of Bi-Functional Molecules by Diversification of the Fatty Acid Pathway. *Metab. Eng.* **2016**, *35*, 9−20.
(535) San, K.-Y.; Wang, D. Alpha Omega Bifunctional Fatty Acids. WO 2016/179572 A1, 10 Nov 2016.
(536) Chia, M.; Schwartz, T. J.; Shanks, B. H.; Dumesic, J. A. Triacetic Acid Lactone as a Potential Biorenewable Platform Chemical. *Green Chem.* **2012**, *14*, 1850−1853.
(537) Jornvall, H.; Persson, B.; Krook, M.; Atrian, S.; Gonzalezduarte, R.; Jeffery, J.; Ghosh, D. Short-Chain Dehydrogenases Reductases (Sdr). *Biochemistry* **1995**, *34*, 6003−6013.
(538) Oppermann, U. C. T.; Filling, C.; Berndt, K. D.; Persson, B.; Benach, J.; Ladenstein, R.; Jornvall, H. Active Site Directed Mutagenesis of 3 Beta/17 Beta-Hydroxysteroid Dehydrogenase Establishes Differential Effects on Short-Chain Dehydrogenase/Reductase Reactions. *Biochemistry* **1997**, *36*, 34−40.
(539) Zha, W. J.; Shao, Z. Y.; Frost, J. W.; Zhao, H. M. Rational Pathway Engineering of Type I Fatty Acid Synthase Allows the Biosynthesis of Triacetic Acid Lactone from D-Glucose in Vivo. *J. Am. Chem. Soc.* **2004**, *126*, 4534−4535.

https://doi.org/10.1021/acs.chemrev.1c00147
Chem. Rev. XXXX, XXX, XXX−XXX

8.3. Sebenta de Bioinformática Molecular

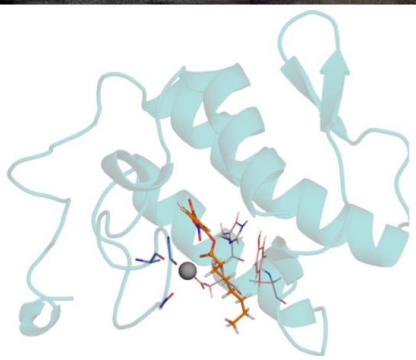
Sousa, J. P.M., Viegas, M. F., Ferreira, P., Paiva, P., Oliveira, A. N.L., Fernandes, P. A., Ramos M. J.

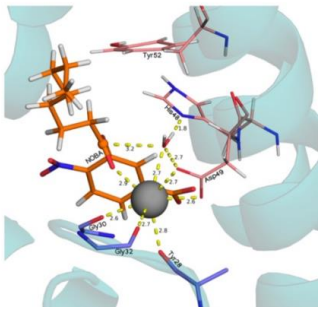
LAQV@REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

BIOINFORMÁTICA MOLECULAR

Conceitos Fundamentais

A QUEDA E A EXPULSÃO DO JARDIM DO ÉDEN por Michelangelo Buonarroti





João Marques Sousa • Matilde Ribeiro Viegas • Pedro Teixeira Ferreira • Pedro Coelho Paiva • Ana Luísa Oliveira • Pedro Alexandrino Fernandes • Maria João Ramos

Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto - 2021

ÍNDICE

PREFÁCIO 3
CAPÍTULO I. MODELAÇÃO POR HOMOLOGIA 4
1.1 Introdução 4
1.2. A técnica 5
1.2.1. Pesquisa de proteínas homólogas e determinação de identidade sequencial 5
1.2.2. Escolha das proteínas homólogas a ser usadas como molde 8
1.2.3. Construção de um modelo tridimensional da proteína-alvo 8
1.2.4. Avaliação da qualidade da estrutura modelada 10
1.3. Conclusão 12

CAPÍTULO II. ANCORAGEM MOLECULAR 13
2.1. Introdução 13
2.2. Ancoragem proteína:ligando 15
2.2.1. Pesquisa 15
2.2.1.1. Pesquisa com recetor rígido 16
2.2.1.2. Pesquisa com recetor flexível 18
2.2.1.3. Moléculas de água a mediar a associação proteína:ligando 19
2.2.2. Pontuação das poses 20
2.2.2.1. Funções de pontuação baseadas em campos de força 22
2.2.2.2. Funções de pontuação empíricas 22
2.2.2.3. Funções de pontuação baseadas em conhecimento estrutural 24
2.2.3. Desempenho da ancoragem molecular proteína:ligando 25
2.3. Conclusão 28

CAPÍTULO III. TRIAGEM VIRTUAL 31
3.1. Introdução 31
3.2. Esquema geral da triagem virtual 35
3.2.1. Preparação da estrutura do recetor 35
3.2.2. Preparação da biblioteca de compostos 38
3.2.3. Seriação dos compostos da biblioteca 41
3.2.4. Visualização e inspeção humana dos compostos melhor seriados 44
3.3. Conclusões 45

CAPÍTULO IV. MÉTODOS HÍBRIDOS QM/MM 46
4.1. A química computacional 46
4.2. Sistemas químicos a estudar 47
4.2.1. Moléculas pequenas e sistemas pequenos 47
4.2.2. Macromoléculas e sistemas de grande dimensão 47
4.3. QM, mecânica quântica 48
4.3.1. A equação de schrödinger 48
4.3.2. O hamiltoniano molecular 49
4.4. MM, mecânica molecular 50
4.5. QM/MM, mecânica quântica/mecânica molecular 52
4.6. Conclusões 53

PREFÁCIO

A Bioinformática funde as duas áreas da Ciência que têm tido os avanços mais desconcertantes do fim do século XX e início do século XXI. Por um lado as nossas sociedades são regidas pelo poder comunicativo e organizativo, sem precedente, que a informática proporciona. Por outro lado, a biologia, ao desvendar muitos dos segredos da vida, permite ao ser humano vencer a doença, prolongar a saúde, estender a vida, e manipular os seres vivos, a um nível que chega a ser assustador, próprio de "deuses" aos olhos de um homem da antiguidade.

O poder da fusão daquelas duas áreas é imenso mas, para ser materializado, elas têm que ser primeiro fundidas na mente do cientista. Ou seja, o mesmo cientista tem de dominar simultaneamente a biologia e a informática para poder revelar o verdadeiro potencial desta fusão.

Este pequeno livro pretende ser um passo simples nesse sentido. Foca-se numa área específica da bioinformática - a bioinformática molecular. Nesta área, o objeto de estudo são moléculas, com todos os seus átomos explicitamente representados, e simulados pelas leis da física, ora de uma forma mais exata, ora de uma forma mais aproximada. Todas as "moléculas" estudadas são parte de seres vivos e são elementos fundamentais da sua fisiologia. São o que chamamos "moléculas biológicas". Pretendemos simula-las recorrendo a poderosas ferramentas informáticas, e com isso obter respostas a problemas biológicos importantes.

O caso concreto que ilustra o estudo são moléculas de veneno de serpente, mais concretamente de víboras. A toxina em estudo é um dos seus componentes principais, a enzima fosfolipase A2. Ao longo do semestre o estudante aprenderá um conjunto de técnicas bioinformáticas que lhe permitirão estudar a toxina, entender o seu modo de ação, e desenvolver antídotos para a mesma, ilustrando o imenso potencial da bioinformática molecular para resolver complexos problemas biológicos, médicos e sociais.

Esperamos com este livro, e com os protocolos experimentais associados, despertar o bioinformático molecular que há dentro de todos vocês.

CAPÍTULO I.
Modelação por Homologia.

1.1 INTRODUÇÃO.

Modelação por homologia consiste numa técnica para construir uma estrutura tridimensional de uma proteína a partir da sequência da proteína e da estrutura tridimensional de outras proteínas que com ela partilhem um ancestral comum.

O conhecimento da estrutura tridimensional de proteínas é de suma importância para a bioquímica, biologia estrutural, biologia funcional, farmacêutica, medicina e biotecnologia. Isto porque a estrutura tridimensional de uma proteína determina a sua função. Ou seja, a sequência aminoacídica determina a estrutura[1] e a estrutura determina a função.

Infelizmente a estrutura de uma proteína é muito difícil de determinar experimentalmente. As duas técnicas experimentais mais usadas para o efeito são a ressonância magnética nuclear (RMN) e a cristalografia de raios-x. A primeira permite obter a estrutura de proteínas em solução, em várias conformações, mas só é aplicável para proteínas relativamente pequenas, com menos de 200 resíduos de aminoácido, o que limita bastante o seu campo de aplicabilidade, já que em janeiro de 2020 ~70 % das proteínas sequenciadas tinham mais do que 200 resíduos de aminoácido. A segunda exige a formação de um cristal da proteína em questão, bastante puro e muito regular, o que pode variar desde trivial a quase impossível, sendo extremamente difícil em muitos casos, em especial se considerarmos proteínas que estão localizadas em meios heterogéneos, tais como proteínas transmembranares, que tendem a desnaturar quando retiradas do seu meio.

Neste contexto, e dada a facilidade de sequenciação de proteínas, não é de admirar que existam muitas mais proteínas sequenciadas do que proteínas com estrutura tridimensional conhecida. No total, em janeiro de 2020, existiam 180 milhões de sequências na base de dados Uniprot, e 80 mil estruturas tridimensionais de proteínas diferentes no banco de dados *Protein Data Bank*, significando que apenas cerca de 0,04 % das proteínas sequenciadas têm estrutura tridimensional conhecida. Estes números mostram que é urgente desenvolver métodos automáticos para determinar a estrutura de proteínas a partir da sequência.

Felizmente as estruturas tridimensionais das proteínas não são tão variáveis quanto a sua sequência, sendo atualmente reconhecidas apenas ~1.300 "formas" tridimensionais diferentes, mais conhecidas como enrolamentos (do inglês *folds*). Ou seja, a diversidade de enrolamentos é muito mais baixa do que a diversidade de sequências. Adicionalmente, verificou-se que proteínas homólogas (ou seja, proteínas que divergiram de um ancestral comum) tendem a conservar muito mais a sua estrutura tridimensional do que a sua sequência. Isto significa que proteínas homólogas têm muitas vezes o mesmo enrolamento e, como tal, estruturas tridimensionais semelhantes, ainda que as suas sequências não sejam tão semelhantes assim. É bastante frequente encontrar proteínas homólogas que partilham o mesmo enrolamento e que têm estruturas tridimensionais bastante próximas, ainda que só conservem 30-40% da sequência em comum. Esta observação deu origem à técnica de

[1] Há algumas exceções, nomeadamente de proteínas que, com o auxílio de chaperones, adquirem estruturas diferentes das "codificadas" na sua sequência.

modelação por homologia – modelar a estrutura de uma proteína a partir da sua sequência e da estrutura conhecida de uma proteína homóloga.

1.2. A TÉCNICA.

A modelação por homologia consiste essencialmente em quatro etapas: pesquisa de proteínas homólogas para servir como moldes (em inglês *templates*) e alinhamento da sequência das proteínas homólogas com a da proteína-alvo (em inglês *target*, a proteína que queremos modelar); construção de um modelo para a estrutura tridimensional da proteína-alvo a partir do molde escolhido; refinamento da estrutura modelada para a proteína-alvo; avaliação da qualidade da estrutura modelada.

Cada um destes estágios vai agora ser descrito em pormenor.

1.2.1. Pesquisa de proteínas homólogas e determinação de identidade sequencial.

Duas proteínas são identificadas como homólogas se elas apresentam semelhanças muito significativas na sua sequência. Sabe-se hoje que essa semelhança significativa só pode ser fruto de ancestralidade comum. No entanto, o contrário não é verdadeiro: o facto de duas proteínas não partilharem significativa semelhança na sua sequência não implica necessariamente a ausência de um ancestral comum. Apenas implica que não se consegue averiguar a ancestralidade comum através de um simples teste de semelhança na sequência.

Mas o que significa exatamente "semelhança significativa" na sequência aminoacídica? Basicamente que a semelhança é muito maior do que seria estatisticamente espectável se ela fosse fruto do mero acaso. Para isso, é preciso tomar em conta o tamanho da base de dados de sequências dentro da qual se procuram homólogos da proteína-alvo. Suponhamos que a base de dados a analisar contém 50 milhões de sequências. A pergunta a fazer é: qual é a maior semelhança que é expectável encontrar entre a sequência da proteína-alvo e a mais parecida dos 50 milhões de sequências, numa perspetiva meramente estatística? Esta semelhança serve como "limite de significância". Se a comparação encontrar uma sequência muito mais semelhante do que a dada pelo "limite de significância", então acreditamos que tal não é fruto do acaso, mas sim uma consequência da partilha de um ancestral comum, e de uma divergência pouco extensa em relação ao ancestral comum.

De facto, a avaliação da semelhança pode ser feita através das sequências de DNA constantes nas bases de dados e da sequência de DNA que codifica a proteína-alvo (DNA:DNA), do DNA traduzido das proteínas da base de dados e da sequência aminoacídica da proteína-alvo (DNA:proteína), ou ainda entre as sequências aminoacídicas das proteínas da base de dados e da proteína-alvo (proteína:proteína). O último caso é o mais preciso. Uma comparação DNA:DNA pode detetar homologia até um limite de divergência de cerca de 200-400 milhões de anos, mas uma comparação proteína:proteína pode detetar homologia até um limite de 2.5 mil milhões de anos (*i.e.* humano-bactéria!).

Ainda que a maior parte dos algoritmos bioinformáticos usados para este fim, tal como o BLAST, SSEARCH, FASTA, ou HMMER, detetem homologia com base na estatística, a maior parte dos investigadores prefere usar outro indicador: a identidade sequencial. Esta consiste na percentagem de resíduos iguais entre a proteína-modelo e a proteína-alvo, com uma linha de corte típica de 30%. Esta linha de corte para a identidade sequencial é "generosa"; de facto a deteção estatística é mais sensível e consegue detetar um maior número de homólogos (cerca de mais 1/3, homólogos remotos) que a identidade sequencial truncada a 30% mas,

como se verá adiante, proteínas homólogas com baixa identidade sequencial são, muitas vezes, pouco úteis para efetuar modelação por homologia.

A identidade sequencial rigorosa é demorada de calcular, porque exige um alinhamento sequencial preciso em toda a extensão das duas sequências, algo que não é feito na pesquisa de homologia baseada na estatística. Ou seja, os *software* BLAST, ou FASTA usam métodos estatísticos aproximados e muito rápidos para detetar homologia por análise estatística de enormes bases de dados, com milhões de sequências. Após identificar um pequeno conjunto de proteínas confirmadamente homólogas, um alinhamento sequencial, rigoroso e demorado, e cálculo da identidade sequencial, pode ser efetuado para uma lista mais reduzida de homólogos mais semelhantes.

O alinhamento sequencial é um processo complexo em que se alinha e compara a sequência de duas ou mais proteínas (neste último caso teremos um alinhamento multisequencial, um processo mais demorado) e se tenta encontrar a melhor correspondência possível entre os resíduos de aminoácido entre elas. Se as proteínas tiverem um número de resíduos aminoacídicos semelhante, e bastantes resíduos em comum, o alinhamento é trivial. Se as proteínas a alinhar tiverem dimensões muito diferentes, e muitas zonas dissemelhantes, o processo é matematicamente mais exigente e não há a garantia de se ter encontrado o melhor alinhamento. Um exemplo de alinhamento sequencial para a fosfolipase A2 secretada de duas víboras crotalíneas.

```
SVVELGKMIIQETGKSPFPSYTSYGCFCGGGERGPPLDATDRCCLAHSCCYDTLPDCSPKTDRYKYKRENGEIIC  75
SLLEFGKMILETGKLAIPSYSSYGCYCGWGGKGTPKDATDRCCFVHDCCYGNLPDCNPKSDRYKYKRVNGAIVC  75

ENSTSCKKRICECDKAVAVCLRKNLNTYNKKYTYYPNFWCKGDIEKC                              122
EKGTSCENRICECDKAAAICFRQNLNTYSKKYMLYPDFLCKGELK--                              120
```

Figura I.1: alinhamento sequencial entre a fosfolipase A2 secretada do veneno da víbora de escamas-em-serra (*Echis Carinatus*, sequência superior) e a víbora de Russell (*Daboia Russelii*, sequência inferior), ambas venenosíssimas. As víboras são mostradas na figura, com a víbora de escamas-em-serra à esquerda e a víbora de Russell à direita. Os resíduos diferentes estão assinalados a cinza-claro e os resíduos em falta estão assinalados com hífens. A identidade sequencial é de 66%.

Outro conceito relevante é a similaridade sequencial. A similaridade sequencial traduz quão semelhantes são duas sequências, através de um sistema de pontuações para cada par de resíduos no alinhamento sequencial. O sistema de pontuações é geralmente calculado através da frequência com que se encontra cada par de resíduos alinhados numa base de dados de proteínas homólogas, dividida pela frequência com que se deveria encontrar o mesmo par numa distribuição aleatória. A razão das duas probabilidades é depois logaritmizada e transformada numa pontuação (*score*).

Se dois resíduos de um par de sequências forem iguais eles contribuem sempre com uma pontuação positiva e elevada. Como a evolução tende a conservar resíduos importantes, e a permitir mais facilmente mutações por resíduos semelhantes do que por resíduos muito diferentes, a pontuação de um par de resíduos acaba por espelhar o grau de semelhança entre eles. Num par de resíduos diferentes, a pontuação pode ser positiva, nula ou negativa, conforme o grau de semelhança entre o par.

O sistema de pontuação BLOSUM62 é um dos mais utilizados, por ser muito versátil e razoavelmente adequado tanto para proteínas com elevado como com reduzido grau de homologia. Neste sistema a conservação de uma leucina tem a pontuação de +4, substituição de uma leucina por uma valina (resíduos muito semelhantes) tem uma pontuação de +1, mas a substituição de uma leucina por uma lisina (resíduos muito diferentes) tem a pontuação de -2. A similaridade global é medida pela pontuação total de todos os pares de aminoácidos alinhados. A matriz de pontuação do sistema BLOSUM62 é mostrada abaixo.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Figura I.2: a matriz de pontuação BLOSUM62, evidenciando a pontuação atribuída a cada par de resíduos num alinhamento sequencial. Num par de sequências aleatórias o valor esperado da pontuação é -0.52. No caso ilustrado na figura I.1, a similaridade sequencial tem um valor elevado de +0.54.

Por fim, saliente-se que nem todas as proteínas homólogas são úteis para fazer modelação por homologia. Para serem úteis, e servirem de molde, as proteínas homólogas têm que ter a estrutura tridimensional determinada. Para simplificar a pesquisa, o *software* de modelação de homologia não faz a pesquisa de homólogos diretamente nas bases de dados de sequências (que contêm neste momento 180 milhões de sequências), mas sim no *Protein Data Bank* (que "só" contém 80 mil estruturas de proteínas únicas), pois estas são as proteínas que têm a estrutura tridimensional conhecida. É frequente os programas de modelação por homologia criarem as suas próprias bases de dados, através do tratamento das entradas encontradas no *Protein Data Bank*, para retirar redundâncias e facilitar a pesquisa.

1.2.2. Escolha das proteínas homólogas a ser usadas como molde.

Após identificar as proteínas homólogas à proteína-alvo, e determinar a sua identidade e similaridade sequenciais, deve proceder-se à escolha das melhores proteínas para serem usadas como molde. O primeiro parâmetro a ser usado é a identidade/similaridade sequencial. Em geral, quanto maior a identidade/similaridade sequencial entre a proteína-alvo e o molde, melhor será o modelo construído. Se a identidade sequencial estiver abaixo de 30%, o modelo é geralmente especulativo. Se a identidade sequencial estiver entre 30% e 60% o modelo é geralmente preciso. Acima de 60%, o modelo construído é geralmente de alta precisão.

Note-se que a exatidão no alinhamento diminui com a diminuição da identidade sequencial, e erros no alinhamento são o fator que provoca maiores erros na construção do modelo. Modelar uma região da proteína-alvo a partir de uma região errada do molde (devido a erros no alinhamento) tem obviamente resultados desastrosos.

Se dois moldes revelarem uma identidade/similaridade sequencial semelhantes, e se as falhas (em inglês *gaps*) no alinhamento ocorrerem na mesma região, então é preferível usar o molde cuja determinação cristalográfica tiver melhor resolução. No entanto, se os dois moldes contiverem falhas no alinhamento em regiões diferentes (o que é bastante comum) então a melhor alternativa é usar ambos os moldes em simultâneo, servindo cada um dos moldes para modelar a proteína-alvo na região onde a identidade/similaridade sequencial local for maior.

Para além destes fatores, há outros aspetos relevantes a considerar. Por exemplo, se a proteína-alvo contiver cofatores metálicos é preferível usar um molde com menor resolução mas que os contenha também do que um molde com maior resolução mas sem cofatores. O mesmo pode ser dito para a presença de ligandos.

Por fim, outros aspetos como o pH, o solvente, ou interações com outras proteínas podem ser importantes na escolha do molde. O melhor molde será aquele que partilhe o mais possível estes fatores ambientais com a proteína-alvo.

No caso da proteína que estamos a modelar foi escolhida a estrutura 1fv0.pdb, correspondente à enzima homóloga da víbora de Russel, cujo alinhamento sequencial foi ilustrado na figura I.1.

1.2.3. construção de um modelo tridimensional da proteína-alvo.

Escolhido o molde (ou os moldes), diversas técnicas podem ser usadas para gerar o modelo tridimensional da proteína-alvo. Os métodos mais comuns são a modelação por montagem de corpos rígidos, a modelação por correspondência de segmentos e a modelação por satisfação de restrições espaciais.

Na modelação por **montagem de corpos rígidos** o modelo é construído através da simples e direta justaposição de porções de proteína retirados do(s) molde(s) escolhido(s), transferindo as coordenadas cartesianas das regiões mais conservadas dos moldes para o modelo. As regiões menos conservadas são geralmente os laços (do inglês *loops*), e são construídas automaticamente, por forma a unir as regiões conservadas. A construção pode ser feita de duas formas: *ab-initio* (ou seja, modelação puramente computacional, resíduo a resíduo, sendo que a conformação dos resíduos modelados é escolhida por forma a otimizar uma função de pontuação de base física ou estatística) ou por correspondência com outros laços semelhantes (ou seja, efetuando uma pesquisa numa base de dados de fragmentos proteicos e selecionando os que têm sequência semelhante às regiões não conservadas, usando a seguir esses fragmentos como moldes).

A **modelação por correspondência de segmentos** tem como princípio o conhecimento parcial da estrutura da proteína alvo (por exemplo, a posição dos carbonos alfa, ou de alguns dos carbonos alfa). Este conhecimento pode provir de várias fontes, quer de conhecimento experimental quer de um bom alinhamento com uma estrutura homóloga. A proteína-alvo é depois dividida em pequenos fragmentos, e uma pesquisa numa base de dados de fragmentos proteicos encontra pequenos "fragmentos correspondentes". Estes fragmentos correspondentes são identificados com base não só na identidade sequencial com os fragmentos da proteína-alvo, mas também na similaridade tridimensional entre as posições e torções dos átomos dos fragmentos correspondentes e dos fragmentos da proteína-alvo. O modelo é então construído importando e justapondo as posições atómicas dos fragmentos. Neste processo são escolhidos os fragmentos correspondentes que melhor evitam colisões estereoquímicas e que minimizam a energia de dispersão e repulsão.

A **modelação por satisfação de restrições espaciais** utiliza métodos matemáticos para satisfazer um conjunto de restrições espaciais a que a estrutura da proteína alvo deve obedecer o mais possível. São geralmente usados dois tipos de restrições: **restrições baseadas em homologia e restrições baseadas em estereoquímica**. Por exemplo, a distância entre dois carbonos alfa da cadeia principal de um dado resíduo pode ser restrita a ser semelhante i) à distância típica para esse tipo de resíduo numa base alargada de proteínas, ii) à conformação do mesmo resíduo em proteínas semelhantes iii) à distância que é observada na região identificada como correspondente no alinhamento com uma proteína homóloga.

Este tipo de restrições é geralmente aplicado não só à distância entre carbonos alfa mas ainda às distâncias entre os átomos de azoto e oxigénio da cadeia principal, aos ângulos diedros da cadeia principal e aos ângulos diedros das cadeias laterais. Outros tipos de restrições, tais como a adoção de valores padrão para comprimentos e ângulos de ligação, são também usados e podem ser extraídos de campos de força de mecânica molecular para proteínas. Após o estabelecimento das restrições, pode-se usar um método de otimização de primeira ordem, e dinâmica molecular, para construir um modelo que viole o mínimo possível a totalidade das restrições.

Em geral, e independentemente do método de modelação, pode-se dizer que a tarefa mais delicada da montagem de qualquer modelo é a construção dos laços, dada a sua grande variabilidade estrutural. Quer sejam construídos *ab initio* quer sejam retirados de fragmentos homólogos, há limites para o comprimento de um laço que pode ser modelado de forma fiável, sendo que se considera seguro modelar laços até oito resíduos (boa exatidão em 90% dos casos). Laços com doze resíduos já só são precisos em 30% dos casos.

Para concluir enfatiza-se que os três métodos de construção de modelos aqui discutidos, quando corretamente usados, conduzem a resultados semelhantes e da mesma qualidade. A escolha das proteínas homólogas que vão servir como molde, e a qualidade do alinhamento, têm muito mais impacto no modelo criado do que o método específico usado para a sua construção.

A figura I.3 abaixo ilustra a sobreposição entre um modelo da fosfolipase A2 secretada do veneno da víbora de escamas-em-serra e a estrutura cristalográfica desta mesma enzima. O modelo usou a enzima homóloga da víbora de Russell como molde, com identidade sequencial de 66%. Como se pode ver, o esqueleto do modelo é praticamente igual ao esqueleto da proteína determinada experimentalmente, sendo as diferenças confinadas aos laços. As cadeias laterais apresentam mais diferenças, mas ainda assim o modelo é de excelente qualidade. O desvio padrão nas posições atómicas (RMSd) para o esqueleto é de cerca de 3.1 Å.

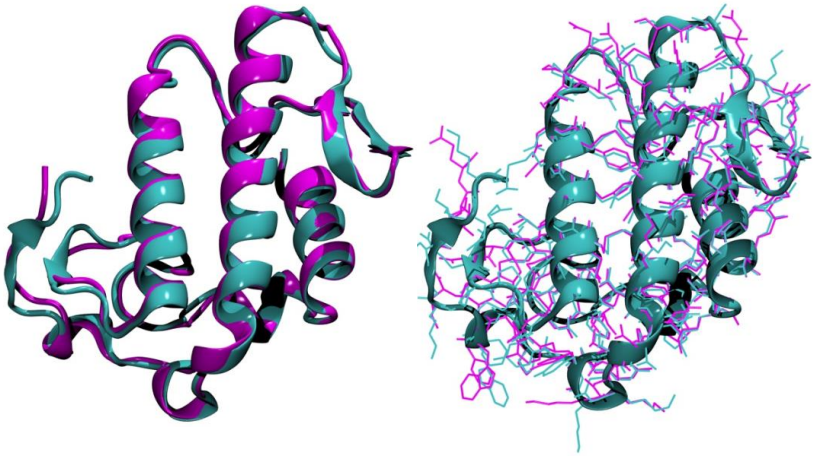


Figura I.3: Esquerda: Sobreposição entre o esqueleto proteico do modelo de homologia da fosfolipase A2 secretada do veneno da víbora de escamas-em-serra (magenta) e a sua estrutura cristalográfica (azul). Direita: As cadeias laterais do model (magenta) sobrepostas ao esqueleto e cadeias laterais da estrutura experimental (azul). O modelo usou como molde a enzima homóloga da víbora de Russell (*Daboia Russelii*), com identidade sequencial de 66%.

1.2.4. Avaliação da qualidade da estrutura modelada.

Uma vez construído o modelo, o passo seguinte é avaliar a qualidade e fiabilidade do modelo gerado. A identidade sequencial, só por si, é já um bom indicador da qualidade do modelo. Modelos construídos a partir de proteína homólogas com mais de 30% de identidade sequencial tendem a ser de boa qualidade, sendo tanto melhores quanto maior a identidade sequencial. No entanto, se a proteína homóloga diferir em outros fatores, para além da sequência, tais como a presença de ligandos, cofatores metálicos, pH, etc., a identidade sequencial, por si só, pode não ser suficiente para se ter uma noção geral da qualidade do modelo.

Existem dois tipos de avaliação do modelo: avaliação interna e avaliação externa. A avaliação interna visa averiguar a auto-consistência do modelo; verifica se os pressupostos de modelação foram cumpridos (por exemplo, se as restrições espaciais foram respeitadas).

A avaliação externa é geralmente bastante mais informativa e usa informação não contida no modelo gerado.

Como avaliação interna é habitual averiguar se os comprimentos de ligação, ângulos de ligação, ângulos diedros, distâncias interatómicas, pontes dissulfureto, diagramas de Ramachandran, etc., se assemelham ao que é tipicamente encontrado em proteínas nativas de alta resolução. Ou seja, avaliam a qualidade estereoquímica do modelo. Algoritmos tais como o Procheck ou Whatcheck ou são comummente usados. Estes algoritmos não detetam "erros" na estrutura. Detetam sim regiões "atípicas", que podem ser fruto de erros de modelação ou não. Seja como for, estas regiões de estrutura atípica devem ser inspecionadas cuidadosamente.

Em relação à avaliação externa, algoritmos como o **ProsaII** ou o **Qmean**, tentam avaliar a estrutura terciária da proteína (o seu enrolamento) no sentido de verificar estatisticamente quão "comum" o seu enrolamento é. Este parâmetro é expresso através de uma pontuação que traduz o "grau de natividade" da estrutura -quanto maior a sua semelhante em relação a proteínas cuja estrutura foi determinada experimentalmente, melhor será a sua pontuação. A pontuação é geralmente calculada com base em potenciais físicos, que calculam dados energéticos sobre o modelo, e comparam esses dados com os obtidos numa base de dados de estruturas experimentais. Zonas de elevada instabilidade poderão indicar a erros de modelação.

Por exemplo, o algoritmo ProsaII calcula a energia da proteína e compara-a com a uma distribuição de energias derivada de um conjunto de conformações aleatórias. É expectável que a energia de uma proteína bem "enrolada" seja mais negativa (mais estável) do que uma estrutura aleatória. Para facilitar a comparação, o algoritmo atribui uma pontuação à energia, e compara-a com a pontuação de estruturas experimentais. No caso da modelo de homologia da fosfolipase A2 secretada do veneno da víbora de escamas-em-serra, essa pontuação é bastante favorável e em linha com a de estruturas nativas experimentais, como se pode ver na figura I.4.

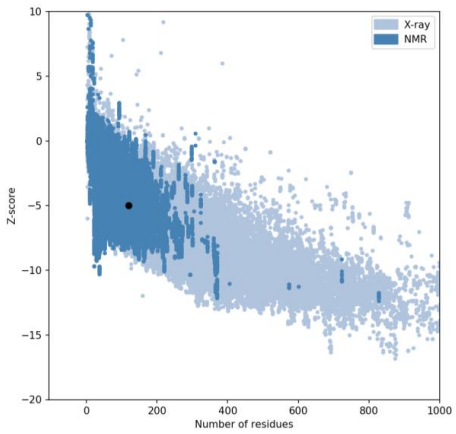


Figura I.4: A pontuação da estrutura modelada acima é de -4,67 (círculo negro), o que está bastante bem enquadrado com o que é expectável de uma estrutura experimental, tal como ilustrado no gráfico pelos círculos

azuis, confirmando a boa qualidade da estrutura tridimensional do modelo. Gráfico retirado do servidor Prosa-Web (*https://prosa.services.came.sbg.ac.at /prosa.php*)

Um aspeto primordial a avaliar é a adequação da escolha da proteína homóloga que foi usada como molde (*i.e.* se a proteína homóloga terá realmente uma estrutura semelhante à proteína-alvo). Quando a identidade sequencial é baixa (<30%), e o alinhamento é incerto, é muito difícil de distinguir uma má modelação feita com um molde correto de uma boa modelação feita com um molde incorreto. Se o molde não for adequado, o modelo será incorreto, mesmo que a modelação tenha sido corretamente efetuada. Ou seja, todo o processo estará errado desde a raiz.

Para avaliar este aspeto costuma-se calcular a pontuação da avaliação da estrutura terciária do modelo e do molde, por exemplo com o algoritmo ProsaII. Pontuações semelhantes costumam indicar a escolha de um modelo adequado embora, mais uma vez, existam exceções a esta regra. No caso do modelo que construímos, a pontuação da estrutura que serviu como molde (ficheiro 1fv0.pdb, enzima homóloga da víbora de Russel), a pontuação era -4,98, o que compara muito bem com a pontuação de -4,67 da estrutura modelada.

1.3. Conclusão.

A modelação por homologia permite gerar estruturas tridimensionais de proteínas a partir da sequência de uma proteína-alvo e da estrutura de uma proteína homóloga. O processo gera estruturas de significativa precisão se a sequência da proteína-alvo e da proteína homóloga forem significativamente semelhantes. Os enormes avanços que a genómica tem tido nas últimas décadas levaram a uma explosão na quantidade de sequências disponíveis, que não tem sido acompanhada por um progresso equivalente na sua determinação estrutural. O fosso entre o número de proteínas com sequência conhecida e com estrutura conhecida tem, portanto, vindo a aumentar. Neste contexto, é expectável que a modelação por homologia assuma um papel de importância crescente, para diminuir as consequências deste fosso no conhecimento. Importa, pois, dominar e desenvolver esta técnica, que previsivelmente continuará em voga durante muitos anos vindouros.

CAPÍTULO II
Ancoragem Molecular

2.1. INTRODUÇÃO.

Ancoragem molecular (do inglês, *molecular docking*) consiste numa técnica que prevê a geometria de associação (também conhecida por pose) de duas moléculas, a partir do conhecimento da estrutura molecular de cada uma delas. A técnica é bastante usada para prever a geometria de associação de complexos proteína: ligando e proteína: proteína, entre outros complexos moleculares.

As associações proteína:ligando são muito comuns na natureza. A proteína tem uma dimensão muito maior que o ligando e tem uma cavidade onde o ligando entra e se liga, na maior parte dos casos de forma não-covalente. As enzimas, por exemplo, são um protótipo deste fenómeno, pois todas têm um centro ativo onde um pequeno ligando (o substrato) se liga, e muitas têm centros alostéreos, onde pequenos ligandos se ancoram, aumentando ou diminuindo a sua atividade catalítica. Alguns têm ainda centros de especificidade, onde a ligação de pequenas moléculas muda a especificidade da enzima para outro substrato, dependendo das necessidades fisiológicas do organismo em cada instante.

Uma grande fração dos fármacos consiste em inibidores de uma dada enzima envolvida na patologia que se quer tratar. Nestes casos, os fármacos são geralmente pequenas moléculas, à imagem do substrato ou do agente alostéreo, que se ligam quer no centro ativo quer nos centros alostéreos da enzima (figura II.1).

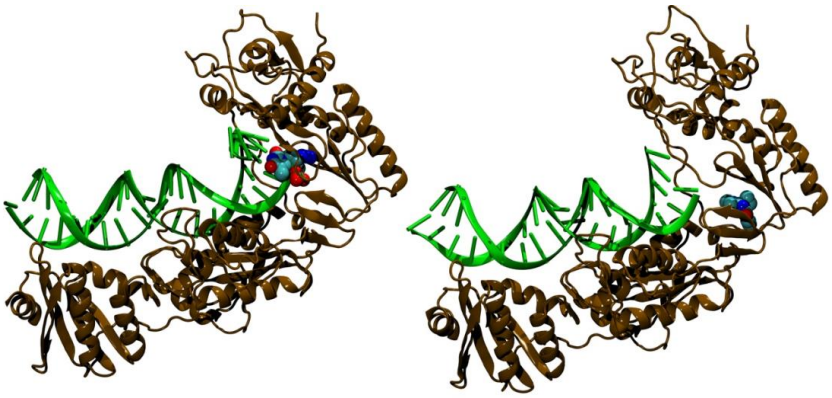


Figura II.1. Esquerda: A enzima transcriptase reversa do vírus HIV-1 (ocre) complexada com um dímero RNA:DNA (verde) e com o inibidor competitivo AZT, ligado no centro ativo (modelo em esferas, colorido por elemento). Direita: A enzima transcriptase reversa do vírus HIV-1 (ocre) complexada com um dímero RNA:DNA (verde) e com o inibidor alostéreo nevirapine, ligado no centro alostéreo (modelo em esferas, colorido por elemento). Os locais de ligação são diferentes, embora próximos. A função da atracagem molecular é prever a pose dos inibidores quando ligados à enzima, partindo das estruturas individuais da enzima e dos respetivos inibidores.

Ainda que seja possível determinar a estrutura da associação proteína:ligando em alguns casos, essencialmente através de cristalografia de raios-x, muitas vezes essa determinação não está disponível, e é de muito difícil e morosa execução, pelo que se torna útil prevê-la através

de ancoragem molecular. No contexto do processo de descoberta de fármacos, é habitual testar a capacidade de ancoragem de milhões de moléculas no centro ativo ou alostéreo de uma dada enzima. Nesses casos, é completamente impossível determinar a estrutura cristalográfica de tal número de compostos, pelo que a única solução técnica para estudar a sua estrutura de associação e prever a sua pose é através da ancoragem molecular. Assim sendo, a ancoragem molecular assume um papel fundamental na bioquímica e biologia estruturais, bem como na descoberta de fármacos.

A ancoragem molecular é também muito útil para prever as geometrias de associação de complexos entre proteínas. Estes complexos proteicos estão envolvidos em fenómenos muito relevantes tais como a sinalização celular, controle do ciclo celular, replicação de DNA, transcrição e tradução de RNA, modificações pós-tradução, entre outros. No entanto, devido à dimensão das proteínas envolvidas nos complexos, e à natureza muitas vezes transiente destes complexos, as suas estruturas são mais difíceis de determinar experimentalmente do que as estruturas de complexos proteína:ligando. Assim sendo, foram também desenvolvidos algoritmos computacionais para prever a pose de um complexo proteína:proteína, partindo da estrutura das proteínas individuais (**Figura II.2**).

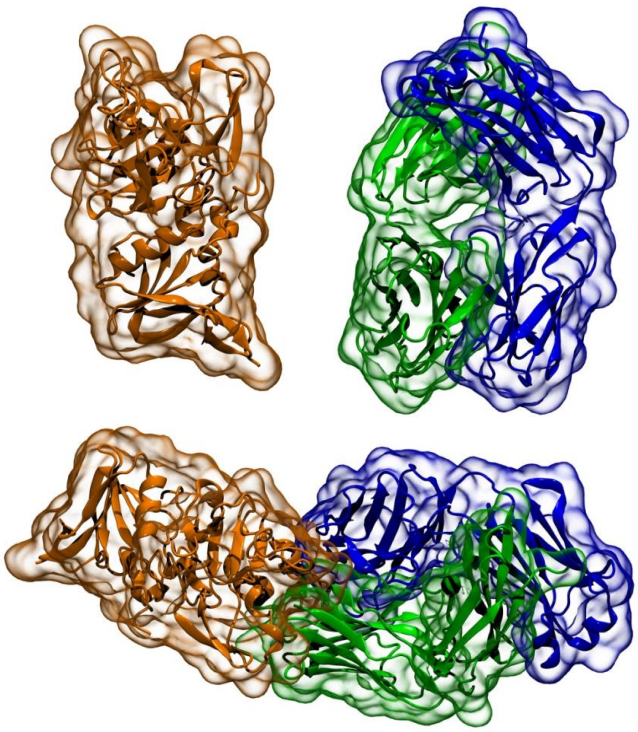


Figura II.2: A figura representa o antigénio AMA1 (a laranja), que é um alvo fundamental para a elaboração de vacinas conta o *Plasmodium Falciparum*, o parasita causativo da malária, e o anticorpo monoclonal 1F9 (fragmento Fab, cadeia leve a verde, cadeia pesada a azul) usado para neutralizar o antigénio AMA1. A atracagem

proteína:proteína determinou a pose do complexo antigénio:anticorpo (topo) a partir da estrutura isolada das proteínas separadas (base).

Estes fenómenos fazem parte de um fenómeno mais vasto, denominado reconhecimento molecular, que consiste na forma como determinados pares de moléculas se "reconhecem" entre si, e se ligam uma à outra, de forma específica e seletiva, através do estabelecimento de interações moleculares, tais como pontes de hidrogénio, interações iónicas, ou interações dispersivas. Por vezes podem estar envolvidas ligações covalentes também.

Em seguida vamos descrever em pormenor as técnicas de ancoragem molecular proteína:ligando e proteína:proteína.

2.2. ANCORAGEM PROTEÍNA:LIGANDO.

A ancoragem molecular computacional ocorre em dois estágios: pesquisa e pontuação. Durante o processo de pesquisa, os algoritmos exploram o máximo de possibilidades para o posicionamento do ligando na proteína (também designada por recetor). No processo de pontuação, o algoritmo atribui uma pontuação a cada pose, de forma a fazer uma seriação e escolher a pose mais provável.

2.2.1. Pesquisa

Existem dois tipos de pesquisa: pesquisa com o recetor rígido e pesquisa com o recetor flexível. Em ambos os casos o ligando é tratado como sendo flexível, e o algoritmo de pesquisa explora, o mais possível, as muitas conformações do ligando que sejam estáveis e que não provoquem colisões com o recetor. Técnicas que consideram o ligando rígido foram usadas no passado, quando o poder computacional era muito mais limitado do que atualmente. O recetor pode ser tratado como tendo uma única conformação (recetor rígido) ou tendo múltiplas conformações (recetor flexível). O primeiro caso vai de encontro ao paradigma de reconhecimento molecular chave-fechadura[2], e o segundo vai de encontro aos paradigmas de reconhecimento molecular de encaixe induzido ou de seleção conformacional. Ainda que a flexibilidade do recetor seja importante (estima-se que 30% das associações proteína:ligando envolvam mudanças conformacionais na proteína) ela é extremamente morosa do ponto de vista computacional, devido ao grande número de graus de liberdade da proteína. Uma estratégia simples para ultrapassar este problema consiste em usar a estrutura de um recetor que contém um ligando no local de ligação, ligando este que já induziu mudanças conformacionais no recetor, e em seguida apagar o ligando e usar esse recetor para ancorar o ligando de interesse, na esperança que as mudanças conformacionais induzidas pelo primeiro sejam semelhantes às induzidas pelo segundo, o que será tanto mais correto quanto mais semelhantes forem os dois ligandos.

O objetivo final do processo de pesquisa é gerar poses o mais próximas possível da pose verdadeira, que é desconhecida, e entre elas garantir que pelo menos uma é muito semelhante à pose verdadeira.

[2] No sentido estrito do termo, o modelo chave-fechadura deveria considerar tanto o recetor como o ligando com sendo rígidos, mas presentemente nenhum software considera o ligando rígido, uma vez que essa aproximação é detrimental para a precisão do resultado.

2.2.1.1. Pesquisa com Recetor Rígido.

Neste modo de pesquisa, apenas a flexibilidade do ligando é explorada. Existem vários algoritmos para analisar a flexibilidade conformacional do ligando (isto é, diferentes conformações do ligando) e a sua colocação no local de ligação (isto é, os graus de liberdade translacionais e rotacionais do ligando como um todo, numa dada conformação). Neste sentido, a geração de poses pode ser decomposta em dois problemas: i) explorar o espaço conformacional do ligando e ii) investigar os seis graus de liberdade (translação e rotação) de cada conformação do ligando no local de ligação.

Uma solução seria a exploração exaustiva: com um método sistemático, rodar cada ligação simples, de 0° a 360°, com incrementos definidos, e criar assim todas as combinações de ângulos diedros para as múltiplas ligações simples do ligando e, por fim, varrer o espaço translacional e rotacional do ligando no local de ligação. O problema da exploração exaustiva é que o número de conformações a explorar cresce muito rapidamente com o número de ligações simples. Considerando incrementos de 15°, o número de conformações de cada ligação simples seria 360°/15°=24. O número de conformações total de um ligando seria 24^N, sendo N o número de ligações simples, o que é intratável, devido à explosão combinatória que a pesquisa sistemática provoca. A figura II.3 ilustra a distribuição de ligações simples em fármacos e em produtos naturais. O valor médio é de 5,5 para os primeiros e 6,6 para os segundos. O limite tipicamente aceite é o de 10 ligações simples para que um composto tenha boa probabilidade de ser biodisponível.

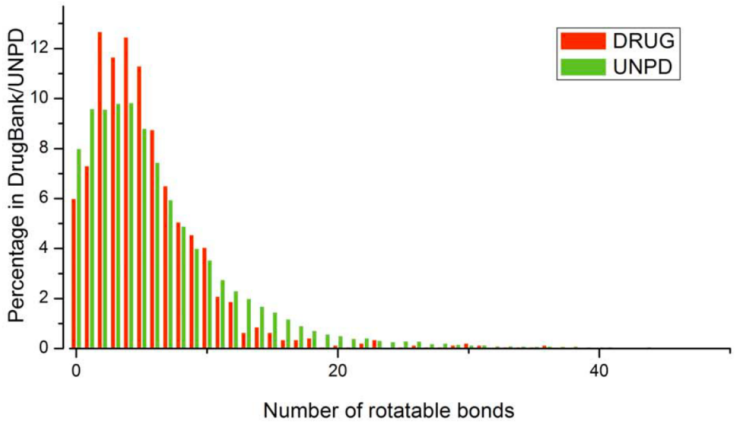


Figura II.3: Distribuição do número de ligações simples (rotáveis) de fármacos (vermelho) e de produtos naturais (verde). Adaptado de *PLoS ONE 8(4):e62839, 2013.*

Com base nestes números, a totalidade de conformações a explorar seria de 39 milhões (5,5 ligações rotáveis), 1286 milhões (6,5 ligações rotáveis), e 63 milhões de milhões (10 ligações rotáveis), o que é naturalmente incomportável, sendo que os métodos sistemáticos só se tornam aplicáveis a moléculas muito pequenas, geralmente sem interesse biológico. Alguns *software* desenvolveram estratégias para limitar o espaço conformacional a pesquisar, de forma a poderem executar uma pesquisa com uma filosofia sistemática mas, ainda assim, limitada e não exaustiva no verdadeiro sentido do termo.

Uma solução para o problema da explosão combinatória na exploração do espaço conformacional consiste no uso de métodos incrementais, também conhecidos como métodos de fragmentação. Nestes métodos o ligando é fragmentado em porções rígidas, cada porção é ancorada no centro ativo através de uma pesquisa do espaço rotacional e translacional e, por fim, as partes flexíveis que ligam os fragmentos rígidos, são montadas de forma a unir as partes rígidas já ancoradas, diminuindo muito as hipóteses conformacionais a explorar. Exemplos de *software* que usam este método são o LUDI, ou o DOCK.

Outra forma de ultrapassar o problema da explosão combinatória de conformações é o uso de métodos estocásticos. Estes métodos geram conformações aleatoriamente e depois aceitam-nas ou rejeitam-nas com base num dado critério, geralmente energético. Um método estocástico bastante usado é o método de Monte Carlo. O método modifica as conformações através de pequenas deslocações aleatórias, e depois aceita-as ou rejeita-as de acordo com a sua probabilidade numa distribuição de Boltzmann. Na sua implementação mais típica, uma nova conformação é aceite de acordo com o algoritmo de Metropolis: se a nova conformação tiver energia mais baixa que a anterior é aceite. Caso contrário, a nova conformação só é aceite se a quantidade $\exp(-\Delta E/K_B T)$ for maior do que um número aleatório sorteado entre 0 e 1 (ΔE é a diferença de energia entre as duas conformações). Estes métodos possibilitam uma pesquisa eficiente das conformações energeticamente mais importantes. Um exemplo de *software* que recorre ao método de Monte Carlo é o ICM.

Por fim, os algoritmos genéticos são também muito usados na fase de pesquisa. Estes algoritmos são inspirados na teoria da seleção natural de Darwin. Cada grau de liberdade de um ligando (por exemplo, o valor de um ângulo diedro, ou uma posição/orientação translacional/rotacional) é codificada num sistema binário, e denominado **gene**; O conjunto dos graus de liberdade de uma pose (os genes) corresponde a um **cromossoma**. Uma função de pontuação avalia a aptidão da pose (do inglês, *fitness*). A exploração das poses prossegue com duas operações genéticas: mutações e cruzamentos. Nas mutações, um grau de liberdade do ligando (um gene) sofre transformações aleatórias. Num cruzamento, um grau de liberdade (um gene) é transferido de uma pose para outra pose. As poses que passam para a etapa seguinte de evolução são aquelas que possuem maior pontuação, ou seja, as mais aptas.

Os *software* AutoDock e GOLD, por exemplo, usam algoritmos genéticos.

A figura II.4 ilustra um processo de ancoragem molecular do voriconazole, um fármaco antifúngico, ligado ao centro ativo da enzima CYP51 de uma amiba altamente patogénica (*naegleria fowleri*, um agente causador de meningoencefalite), com ênfase no processo de pesquisa, ilustrando as diversas poses encontradas e na semelhança entre a pose prevista e a pose cristalográfica verificada no ficheiro 6AY6, depositado no *Protein Data Bank*. A solução prevista pela ancoragem molecular e a estrutura cristalográfica têm um desvio médio (RMSd) de 1.3 Å.

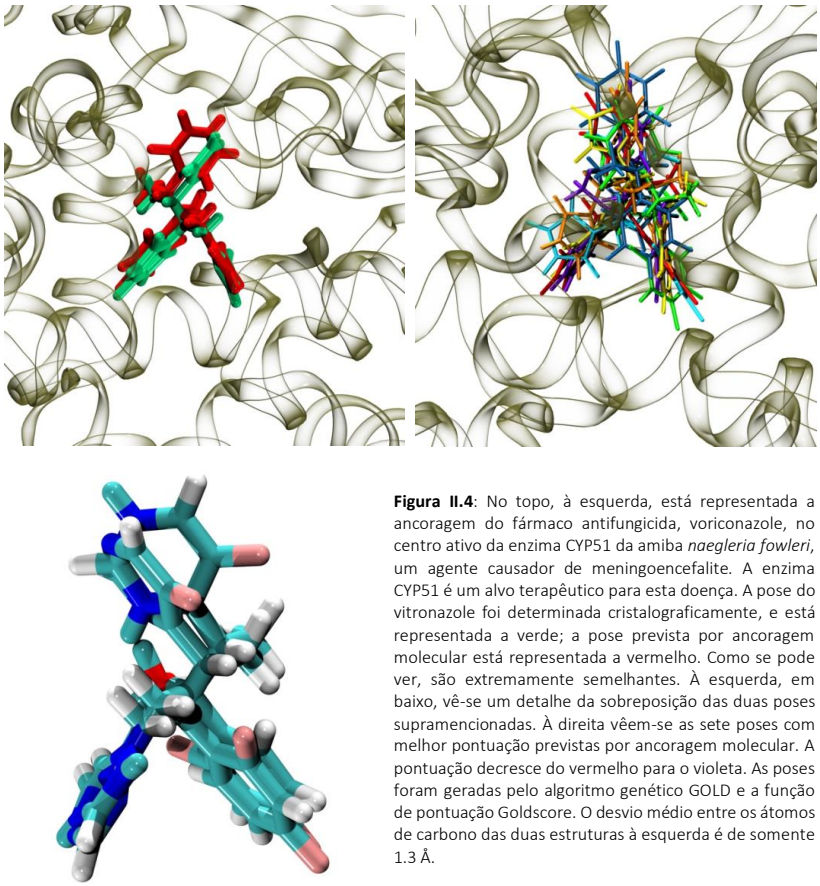


Figura II.4: No topo, à esquerda, está representada a ancoragem do fármaco antifungicida, voriconazole, no centro ativo da enzima CYP51 da amiba *naegleria fowleri*, um agente causador de meningoencefalite. A enzima CYP51 é um alvo terapêutico para esta doença. A pose do vitronazole foi determinada cristalograficamente, e está representada a verde; a pose prevista por ancoragem molecular está representada a vermelho. Como se pode ver, são extremamente semelhantes. À esquerda, em baixo, vê-se um detalhe da sobreposição das duas poses supramencionadas. À direita vêem-se as sete poses com melhor pontuação previstas por ancoragem molecular. A pontuação decresce do vermelho para o violeta. As poses foram geradas pelo algoritmo genético GOLD e a função de pontuação Goldscore. O desvio médio entre os átomos de carbono das duas estruturas à esquerda é de somente 1.3 Å.

2.2.1.2. Pesquisa com Recetor Flexível.

Neste caso, os métodos para explorar a flexibilidade do ligando são semelhantes mas, adicionalmente, também se explora o espaço conformacional do recetor. Este último é muito mais vasto do que o do ligando, em virtude da grande diferença de dimensão entre as duas moléculas. Neste caso, a exploração exaustiva seria inimaginável! Afinal uma proteína com 300 resíduos pode facilmente ter mais de 2,000 ligações simples, e, portanto, o número de conformações a explorar seria 24^{2000}, o que é vastamente superior ao número de átomos no universo conhecido. Como tal, são necessários métodos simplificados para introduzir alguma flexibilidade no recetor. Existem duas categorias de métodos. O primeiro trata algumas cadeias laterais da proteína com algoritmos semelhantes aos que usa para explora a flexibilidade do ligando. Ou seja, expande a pesquisa conformacional de modo a incluir uma porção da proteína, na qual os ângulos diedros das cadeias laterias são completamente livres de se rearranjar, ou são livres de se rearranjar dentro de limites pré-estabelecidos.. Estes métodos

reproduzem melhor o paradigma do encaixe induzido, no qual o rearranjo da enzima é um produto da interação entre a proteína e o ligando.

A segunda família de métodos gera diversas conformações da proteína e depois efetua a ancoragem molecular a cada uma delas. Uma ancoragem deste tipo denomina-se como ancoragem em ensemble. Esta família de métodos reproduz melhor o paradigma de reconhecimento molecular de seleção conformacional, que considera que os rearranjos conformacionais da proteína não resultam da interação direta com o ligando. Considera sim que a proteína já existe num conjunto de conformações que inclui a conformação "rearranjada" que se observa no complexo proteína:ligando. O ligando, por ter mais afinidade para esta, liga-se a ela predominantemente, fazendo baixar a sua abundância. Pelo princípio de Le Chatelier, o equilíbrio conformacional é deslocado no sentido de enriquecer o ensemble na conformação a que o ligando se liga, dando origem a mais associações proteína:ligando, e assim sucessivamente.

Os métodos para gerar as diversas conformações da proteína são, geralmente, de três tipos: potenciais suavizados, bibliotecas de rotâmeros e simulações moleculares.

O método dos **potenciais suavizados** é o mais simples e rudimentar, e consiste em modificar a repulsão interatómica de forma a torna-la mais suave, permitindo alguma sobreposição entre os átomos do recetor e do ligando, representando de forma implícita a flexibilidade do recetor. Ou seja, mimetiza a possibilidade de o recetor se rearranjar para evitar colisões com o ligando.

No caso do uso de **bibliotecas de rotâmeros**, o utilizador decide *a priori* quais as cadeias laterais que quer flexibilizar (geralmente menos que uma dezena) e o algoritmo escolhe as conformações mais estáveis de cada cadeia lateral (os rotâmeros) a partir de uma base de dados, e combina-os entre si de forma a gerar um conjunto de combinações de rotâmeros de baixa energia e sem colisões. Ou seja, em vez de se ter uma conformação para o recetor, tem-se um conjunto de conformações para o recetor, em cada uma das quais se repetem os procedimentos descritos em **2.2.1.1.**

Os métodos baseados em **simulações moleculares** geralmente recorrem à dinâmica molecular. Executa-se uma dinâmica molecular longa do recetor e grava-se um ficheiro de trajetória. Procede-se em seguida a um agrupamento das conformações do recetor com base em similaridade (em inglês, *clustering*), escolhe-se uma estrutura representativa de cada grupo, e usam-se cada uma dessas estruturas para efetuar a ancoragem do ligando. O número de conformações do recetor pode ser escolhido definindo o número de grupos de estruturas resultantes do agrupamento por similaridade. A ancoragem do ligando é efetuada tal como descrito em **2.2.1.1.** em cada uma das estruturas do recetor. Este processo garante que as conformações do recetor retiradas da dinâmica molecular são o mais diversas possível entre si, garantindo uma boa cobertura do espaço conformacional do recetor.

2.2.1.3. Moléculas de água a mediar a associação proteína:ligando.

Um dos principais desafios na ancoragem molecular é a correta previsão de moléculas de água que auxiliam a associação entre o ligando e a proteína. Estas moléculas estabelecem pontes de hidrogénio com o ligando e com a proteína simultaneamente, e podem ser consideradas como uma parte da estrutura da proteína.

Na ausência do ligando durante o processo de cristalização, estas moléculas de água costumam evaporar, ou não ficar fixas no local de ligação, tornando-se indetetáveis. Um

exemplo bastante conhecido é a protease do HIV-1, que liga uma molécula de água juntamente com os seus inibidores, fazendo esta molécula duas pontes de hidrogénio com o inibidor e outras duas com a proteína (figura II.5).
Para superar este problema, é habitual observar estruturas cristalográficas do mesmo recetor mas com outros ligandos (se existirem), para inferir sobre a possível localização das ditas moléculas de água. Pode-se ainda usar *software* apropriado (como por exemplo o programa JAWS) para auxiliar à previsão da existência de moléculas de água a mediar os contactos proteína:ligando. Por fim, simulações de dinâmica molecular feitas *a posteriori* podem (por vezes) recuperar algumas dessas moléculas de água. No entanto, nenhum dos procedimentos é completamente satisfatório, e a previsão da existência de moléculas de água a mediar os contactos proteína:ligando, e a sua inclusão num posição aceitavelmente correta, continua a ser um problema que está longe de ser resolvido.

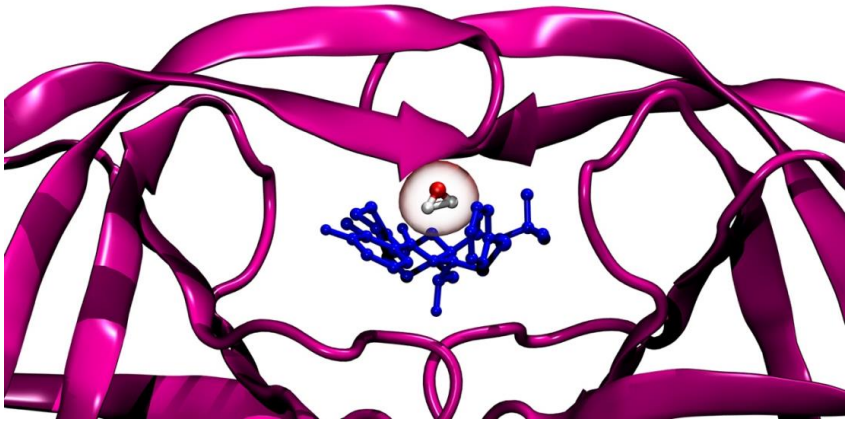


Figura II.5: A figura ilustra uma molécula de água que medeia o contacto entre a protease do HIV-1 e os seus inibidores, e é fundamental para a associação enzima:inibidor (estrutura 1OHR, *Protein Data Bank*). Se a molécula de água não estiver presente na estrutura cristalográfica da enzima, pode tornar-se bastante difícil antecipar que, juntamente com o inibidor, uma molécula de água tem de ser incluída, e exatamente na posição ilustrada.

2.2.2.Pontuação das poses.

Após gerar um conjunto de poses o tão completo e diverso quanto possível, e idealmente contendo um número substancial de poses bastante próximas da pose verdadeira, o passo seguinte consiste em atribuir uma pontuação às poses, por forma a identificar a dominante, *i.e.* a pose observada experimentalmente.

Devido à energia térmica, num sistema real à temperatura ambiente, o complexo proteína:ligando consiste num conjunto muito grande de poses (um ensemble de poses), e não numa pose única. Neste contexto, a abundância de cada pose específica depende da energia do complexo proteína:ligando nessa mesma pose. No entanto, em sistemas proteína:ligando, e devido ao fenómeno do reconhecimento molecular, uma das poses costuma ser muitíssimo mais frequente e abundante do que quaisquer outras, pois nessa dita pose a complementaridade morfológica e eletrostática entre a proteína e o ligando é muito perfeita (tal como previsto no modelo chave-fechadura, ou encaixe induzido, ou seleção

conformacional). Apenas se verificam pequenas oscilações translacionais/rotacionais em torno da pose dominante (geralmente inferiores a 2 Å), e é esta pose que se pretende identificar. É claro que esta observação não é universal, mas é extremamente comum em sistemas proteína:ligando, devido à alta especificidade das interações que o constituem. Assim sendo, pode-se falar numa "pose verdadeira" ou "pose experimental" sem ser excessivamente incorreto. O facto de as estruturas cristalográficas determinarem o ligando numa posição clara e específica constitui uma prova experimental disso mesmo.

Neste contexto, as funções de pontuação visam efetuar uma seriação das poses e escolher a pose mais favorável – a de melhor pontuação - que é identificada como a pose verdadeira. A pontuação reflete geralmente a intensidade das interações atrativas proteína:ligando, a energia necessária para o ligando (e a proteína, no caso de ancoragem com recetor rígido) adquirir a conformação em que foi ancorado, e o custo da dessolvatação da proteína e do ligando, que é ligeiramente diferente em cada pose.

Por vezes a pontuação representa a energia de Gibbs de associação, que é a quantidade mais relevante para traduzir a afinidade proteína:ligando à temperatura ambiente. Considerando a reação de associação:

$$Prot_{(aq)} + Lig_{(aq)} \rightleftharpoons Prot{:}\,Lig_{(aq)}$$

A constante de equilíbrio para a associação, $K_{ass,}$ quantifica a extensão da associação proteína:ligando:

$$K_{ass} = \frac{[Prot{:}\,Lig_{(aq)}]}{[Prot_{(aq)}][Lig_{(aq)}]}$$

E a energia de Gibbs padrão de associação é dada por:

$$\Delta G^0_{ass} = -RTln\left(\frac{K_{ass}}{C_0}\right)$$

Onde C_0 representa a concentração-padrão usada para determinar a energia de Gibbs de associação, que corresponde a 1M em fase condensada[3]. Ainda que seja a quantidade fundamental, a energia de Gibbs de associação é muito complexa de determinar, e as funções de pontuação que a calculam apenas o fazem de uma forma aproximada. Em alguns casos as funções de pontuação obtêm apenas um "índice de afinidade", um valor numérico que a representa, e que permite identificar a pose com maior afinidade, sem que esses valores de pontuação correspondam a uma quantidade física clara e inequivocamente definida.

Existem diferentes funções de pontuação (*scoring functions*), dos quais as mais comuns são as baseadas em campos de forças (do inglês, *force field scoring functions*), empíricas (do inglês, *empirical scoring functions*), e baseadas em conhecimento estrutural (do inglês, *knowledge-based scoring functions*).

[3] O valor de ΔG^0_{ass} para substratos de enzimas ronda os -4 a -6 kcal/mol, e para inibidores clínicos (fármacos) os -8 a -12 kcal/mol.

2.2.2.1. Funções de pontuação baseadas em campos de força.

As funções de pontuação baseadas em campos de força determinam a energia de interação proteína:ligando subdividida em termos intramoleculares (energia interna do ligando devido à conformação adotada) e intermoleculares (energia de dispersão-repulsão e energia eletrostática). A característica fundamental destas funções de pontuação é que a pontuação total depende da soma direta de cada um dos termos, sem que o peso relativo de cada termo na soma tenha sido otimizado por regressão, à imagem da energia calculada por campos de forças tradicionais.

A equação abaixo exemplifica o cálculo da energia baseado numa equação deste tipo, nomeadamente a função **Dock**:

$$E = \sum_{pares\ i,j} 4\,\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{pares\ i,j} \frac{1}{4\pi\varepsilon_{ij}}\frac{q_i q_j}{r_{ij}}$$

Onde os termos a azul são parâmetros pré-tabelados (para a proteína) ou calculados de forma automática (para cada ligando específico). O primeiro termo da equação representa o somatório das interações dispersivas e da repulsão de Pauli entre todos os pares de átomos do sistema separados por mais de três ligações químicas e o segundo representa as interações eletrostáticas para esses mesmos pares de átomos. A solvatação é considerada implicitamente, através da constante dielétrica dependente da distância ε_{ij}. Os parâmetros são retirados do campo de forças do *software* Amber. A pontuação representa aqui a energia de interação.

A exatidão dos campos de força da mecânica molecular é bastante satisfatória mas a sua aplicação é computacionalmente morosa. As funções de pontuação Goldscore, ou GBVI/WSA, entre outras, são também baseadas na filosofia dos campos de forças. Por exemplo, a função Goldscore é dada por:

$$S = S_{hb_ext} + S_{vdW_ext} + S_{hb_int} + S_{vdW_int}$$

Onde S representa a pontuação da pose (do inglês, *Score*). O primeiro termo representa a interação devida a pontes de hidrogénio proteína-ligando; o segundo a interação devida às interações de van der Waals proteína–ligando; o terceiro as pontes de hidrogénio internas do ligando; o quarto representa a tensão conformacional do ligando. Cada um dos quatro termos é calculado a partir de uma função complexa, que envolve parâmetros tais como o tipo de átomo e a distância interatómica, entre outras, que nos escusamos aqui de descrever.

2.2.2.2. Funções de pontuação empíricas.

As funções de pontuação empíricas envolvem termos relativamente semelhantes aos dos campos de forças. A grande diferença é que os termos são combinados linearmente através de coeficientes ajustáveis, cujos valores são otimizados por forma a reproduzir, o melhor possível, um conjunto de afinidades experimentais de uma base de dados de complexos proteína:ligando.

Ou seja, estas funções têm a forma genérica:

$$S = \sum w_i S_i$$

Onde S representa a pontuação total, S_i a pontuação dos diversos termos energéticos e w_i (do inglês, *weight*) o peso com que cada termo é contabilizado para a pontuação total. Funções de pontuação deste tipo incluem o GlideScore e o ChemScore.

A última é dada pela seguinte equação:

$$\Delta G_{bind} = \Delta G_0 + \Delta G_{hbond} S_{hbond} + \Delta G_{metal} S_{metal} + \Delta G_{lipo} S_{lipo} + \Delta G_{rot} S_{rot}$$

Onde S_{hbond}, S_{metal}, S_{lipo} correspondem às pontuações relativas a interações do tipo ponte de hidrogénio, covalente com metais, e dispersiva (lipofílica). O termo S_{rot} representa a penalização entrópica do ligando devido à diminuição da sua liberdade conformacional após a associação. Os termos ΔG_x representam o peso de cada componente de pontuação S_x para a pontuação total ($\Delta G_x G_{bind}$), e resultam de uma regressão multilinear que ajustou a pontuação à energia de Gibbs experimental de associação de 82 complexos proteína:ligando. Neste sentido, a pontuação representa aqui uma energia de Gibbs de associação.

Algumas funções de pontuação têm características híbridas. Cada termo da pontuação deriva de um campo de forças, mas o seu peso é ajustado por regressão. Um exemplo disso é a função de pontuação AutoDock4:

$$S = W_{vdW} \sum_{pares\ ij} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + W_{hb} \sum_{Hb} E(\theta) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + W_{elec} \sum_{pares\ ij} \frac{1}{\varepsilon_{ij}} \frac{q_i q_j}{r_{ij}} +$$

$$+ W_{sol} \sum_{pares\ ij} (S_i V_j + S_j V_i) e^{-r_{ij}^2 / 2\sigma^2}$$

Onde cada termo representa uma interação física, com uma forma matemática similar à usada num campo de forças (energia de van der Waals, energia de pontes de hidrogénio, energia eletrostática e solvatação, respetivamente), e os termos W_x representam o peso de cada termo na pontuação final, e foram ajustados por multi-regressão linear com base na afinidade experimental de 188 complexos proteína-ligando. A energia de Gibbs de associação final, que é o que define a pose mais provável, obtém-se aplicando a equação acima ao complexo proteína:ligando, ao ligando e à proteína, e subtraindo a energia dos monómeros à energia do complexo. Por fim, adiciona-se um termo relativo à perda de entropia do ligando após complexação.

Dado que as funções de pontuação empíricas são ajustadas para reproduzir afinidades (isto é, constantes de associação, ou energias de Gibbs de associação), elas são especialmente aptas para seriar conjuntos de ligandos do mesmo recetor, identificando os ligandos de melhor afinidade. No entanto, foi já demostrado que a afinidade correta só se consegue se for usada a pose correta e, assim, estas funções de pontuação também estão implicitamente otimizadas para identificar a pose correta de entre um grande número de poses.

2.2.2.3. Funções de pontuação baseadas em conhecimento estrutural.

A terceira categoria de funções de pontuação não se baseia em afinidades experimentais, mas sim em bases de dados de estruturas cristalográficas proteína:ligando de alta resolução. A distribuição de Boltzmann diz-nos que a probabilidade de um estado está relacionado com a sua energia, sendo que um dado estado é tanto mais provável quanto menor for a sua energia. Esse é o principio subjacente às funções de pontuação aqui discutidas: assume-se que quanto mais estabilizadora for uma dada interação interatómica para a formação do complexo proteína:ligando, mais frequentemente essa interação deve ser encontrada nas bases de dados cristalográficos. Desta forma, foram criados sistemas de pontuação para interações átomo-átomo, apenas dependentes da distância, com base na frequência com que as interações constam na base de dados.

Estes sistemas de pontuação são muito rápidos de computar - as bases de dados, os átomos são primeiro categorizados em "tipos"[4]. Para cada par de tipos de átomo é contabilizada a frequência do seu contacto. Essa frequência é contabilizada em função da distância interatómica. Quanto mais alta for a frequência com que se encontra um dado par de tipos de átomos a uma dada distância, mais alta será a pontuação para essa configuração estrutural.

Ao focar-se na estrutura dos complexos (e não na afinidade) as funções baseadas em conhecimento estrutural são otimizadas para reproduzir corretamente poses, e não afinidades. Quando o conhecimento da pose é o mais importante estas funções de pontuação são especialmente adequadas.

Um exemplo de uma função deste tipo é o DrugScore:

$$pontuação_{par}(p,l,r) = -\ln\left(\frac{P(p,l,r)}{P_{ref}}\right)$$

Onde p representa um átomo da proteína, l um átomo do ligando, e *r* a distância entre eles. *P(p,l,r)* representa a probabilidade de encontrar o referido par de átomos à distância *r* e P_{ref} corresponde a uma probabilidade de referência, um fator de normalização, que corresponde à probabilidade para o par *p,l* num estado em que os átomos têm uma interação que é a média da encontrada na base de dados.

A probabilidade *P(p,l,r)* é dada por:

$$P(p,l,r) = \frac{N_{p,l,r}}{\Delta V(r)} \cdot \frac{1}{\sum_{r'} N(p,l,r')/\Delta V(r')}$$

onde $N_{p,l,r}$ corresponde ao número de contactos encontrados na base de dados entre os átomos *p* e *l* à distância *r* (mais exatamente entre *r* e *r+dr*), $\Delta V(r)$ é um fator de normalização para o volume da coroa esférica com raios r e r+dr (quanto maior a coroa maior o número de contactos expectáveis), e a fração da direita contabiliza a distribuição dos contactos entre *p* e *l* em função de *r*, e é usada como fator de normalização de $P(p,l,r)$ em relação a *r*.

[4] Um tipo de átomo é uma especificação que, para além do elemento, define o ambiente químico do átomo – a sua vizinhança, carga e hibridização. Por exemplo, um átomo de oxigénio é categorizado em tipos diferentes conforme participe num grupo hidroxilo, éter, carboxilo, ou carbonilo, por exemplo.

Resumidamente, quanto mais vezes se encontrar na base de dados cristalográficos um dado par de tipos de átomos a uma dada distância, maior será a sua pontuação se esse mesmo par de tipos de átomos for encontrado a essa mesma distância nas poses que estão a ser avaliadas.

2.2.3. Desempenho da Ancoragem Molecular Proteína:Ligando.

Os dois estágios do processo de ancoragem molecular costumam ser avaliados independentemente. O poder de pesquisa representa a capacidade de um algoritmo/*software* para gerar a pose verdadeira, e para a identificar entre todas as poses geradas. O poder de seriação representa a capacidade de ordenar um conjunto de ligandos com base na sua afinidade para o alvo. Ou seja, representa a capacidade de prever a afinidade relativa de diferentes ligandos, uma vez que a afinidade absoluta não é prevista pelas funções de pontuação.

O **poder de pesquisa** é influenciado pela eficácia do algoritmo de pesquisa e pela função de pontuação. O primeiro tem de gerar um conjunto de poses que contenha a pose verdadeira, e a segunda tem de a reconhecer entre todas as poses geradas. Assim sendo, costuma-se separar o poder de pesquisa em duas partes: i) a capacidade de gerar a pose verdadeira, e ii) a capacidade de a identificar, sendo que a primeira depende mais do algoritmo de pesquisa, e a segunda da função de pontuação.

É claro que as poses geradas nunca são exatamente iguais à verdadeira, e a pose verdadeira "exata" nunca é conhecida, devido à incerteza experimental. Neste contexto, considera-se uma pose gerada computacionalmente como idêntica à verdadeira se o desvio médio (RMSd) entre os átomos dos ligandos nas duas poses for inferior a 2 Å.

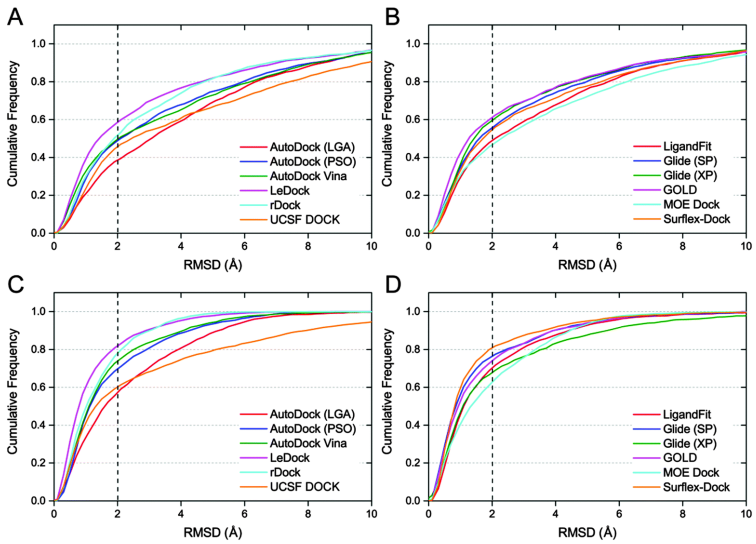


Figura II.6: Painéis A e B: RMSd entre a pose melhor pontuada e a pose experimental, para cinco software de distribuição livre (A) e cinco software comerciais (B). Painéis C e D. RMSd entre a pose mais semelhante à pose experimental, e a pose experimental, para cinco software de distribuição livre (C) e cinco software comerciais (D). Foram usados 2002 complexos proteína:ligando, e para cada complexo foram geradas, no máximo, 20 poses. Figura adaptada de *Phys.Chem.Chem.Phys., 2016, 18, 12964*.

Um estudo recente testou a capacidade de dez programas de ancoragem molecular para reproduzir a pose experimental dos ligandos de um conjunto de 2002 complexos proteína:ligando cuja estrutura experimental é conhecida. Numa primeira análise (**Figura II.6**) calculou-se o RMSd entre a pose melhor pontuada e a pose experimental, e numa segunda análise calculou-se o RMSd entre a pose mais semelhante à experimental (mesmo que não tivesse sido seriada como a melhor pontuada pela função de pontuação) e a pose experimental. Ou seja, a primeira análise reflete a capacidade de gerar e identificar a pose experimental, ao passo que a segunda reflete apenas a capacidade de gerar a pose experimental.

A **Figura II.7** sumariza estes dados, dividindo as poses entre "correta" e "incorreta", sendo que uma pose correta é simplesmente uma pose cujo RMSd para a pose experimental é inferior a 2 Å. A análise foi efetuada quer para a pose melhor pontuada (*top scoring pose*, na figura II.7) como para a pose gerada que era mais semelhante à experimental (*best pose*, na figura II.7). A taxa de sucesso representa a percentagem de estruturas consideradas "corretas". Como se pode ver, a percentagem de sucesso ronda os 40-60%. Em relação à capacidade de gerar a pose correta, o sucesso ronda os 55-80%, mostrando que as funções de pontuação, ainda que demostrem um desempeno absolutamente notável, ainda apresentam deficiências na identificação da pose correta.

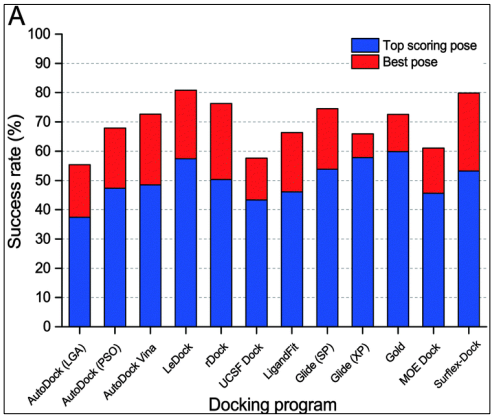


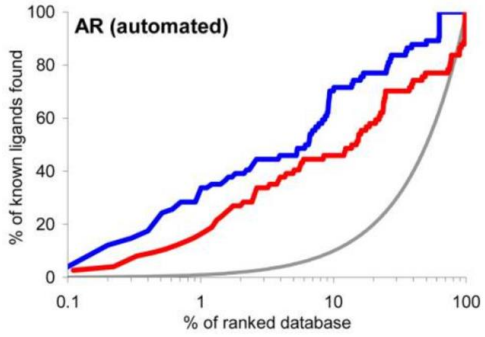
Figura II.7: Percentagem de sucesso na identificação da pose verdadeira entre as poses melhor pontuadas (azul) e na geração das poses verdadeiras (vermelho). Uma pose é considerada verdadeira se a sua diferença (RMSd) para a pose experimental estiver abaixo de 2 Å. Figura adaptada de Phys.Chem.Chem.Phys., 2016, 18, 12964.

Os testes ilustrados nas figuras II.6 e II.7 foram efetuados de forma "automatizada",sem conhecimento bioquímico do utilizador. É claro que num caso real, o utilizador possui conhecimento bioquímico sobre o sistema que estuda (interações preservadas, orientações necessárias para que uma reação ocorra, a natureza química do ligando, etc.) que lhe permitem escolher, de forma mais consciente, quer as poses bioquimicamente consistentes quer as funções de pontuação mais adequadas a cada caso, o que faz com que a percentagem de sucesso seja superior à ilustrada nas figuras II.6 e II.7, que representam testes cegos sobre milhares de complexos. Em casos reais, a percentagem de sucesso na identificação da pose verdadeira ronda os 70-75%.

O **poder de seriação** é relevante para comparar a afinidade de diferentes ligandos. Representa a capacidade de os ordenar corretamente, de acordo com a sua afinidade, e não a capacidade de reproduzir o valor experimental da energia de Gibbs de associação. É claro que a ordenação de diferentes ligandos não depende só da função de pontuação, mas também da capacidade de gerar a pose correta. No entanto, os algoritmos de quase todos os *software*

mostraram uma capacidade bastante satisfatória (e relativamente similar) de gerar poses nativas (55-80% de forma "automática"). Assim sendo, o poder de seriação depende muito mais da função de pontuação do que do método de pesquisa. Existem duas formas comuns de aferir o poder de seriação. A primeira consiste em medir o sucesso com o qual um *software* consegue reproduzir a seriação de um conjunto de ligandos de acordo com a sua afinidade experimental. No entanto, este método é pouco usado porque para cada recetor não costuma haver um grande número de ligandos de boa afinidade e com medição experimental rigorosa. Além disso, a reprodução da seriação é uma tarefa muito difícil, e pouco adequada, para a técnica de ancoragem molecular. Existem outras técnicas computacionais para o fazer, bastante mais exatas mas também mais morosas, que caem na categoria dos "métodos alquímicos", tais como a **integração termodinâmica**, ou a **perturbação de energia livre**, que não serão abordados neste capítulo.

A segunda forma de aferir o poder de seriação consiste em calcular o fator de enriquecimento conseguido numa grande base de dados de ligandos. O fator de enriquecimento é uma medida da capacidade de distinguir ligandos que ligam ao recetor de moléculas que não se ligam ao recetor (não-ligandos). Para tal, constrói-se uma base de dados de moléculas, onde existe um número conhecido de ligandos e de não-ligandos (por exemplo, 20 ligandos e 20.000 não-ligandos). Em seguida efetua-se a ancoragem de todas as moléculas ao recetor e calcula-se a pontuação de cada uma, ordenando-as por afinidade decrescente. Se o processo de ancoragem fosse perfeito, as 20 moléculas melhor pontuadas seriam os 20 ligandos, e só depois apareceriam os não-ligandos. No entanto, tal não acontece, devido às insuficiências dos métodos de ancoragem.



A fração de ligandos nesta base de dados seria de 20/20.000=0,001. Como tal, a probabilidade de identificar um ligando aleatoriamente seria 0,01. Para calcular o fator de enriquecimento define-se primeiro uma fração da lista de seriação (por exemplo, os 1% de compostos melhor pontuados, 200 compostos) e calcula-se a razão entre ligandos e não ligandos entre esses 200 compostos. Se lá existirem 10 ligandos, a razão será 10/200=0,02, o que está muito acima da probabilidade aleatória (0,001). O fator de enriquecimento é a razão entre as duas quantidades, 0,02/0,001 = 20. Ou seja, é vinte vezes mais provável encontrar ligandos em 200 compostos ordenados pela pontuação de ancoragem do que entre 200 compostos selecionados aleatoriamente

Figura II.8: Capacidade de deteção de ligandos do receptor de androgénio humano com o software DOCK 3.5.54. A linha vermelha ilustra os resultados de uma experiência com 74 ligandos e 2.630 não-ligandos muito semelhantes aos ligandos (denominados réplicas). A linha a azul representa uma experiência com 74 ligandos, 2.630 réplicas e 95.636 não-ligandos. A linha azul representa portanto um teste menos complexo que o representado pela linha vermelha. A linha cinza representa a probabilidade de deteção aleatória de ligandos. O fator de enriquecimento a uma dada fração da base de dados é dado pela razão entre valores marcados pela linha azul/verde e linha cinzenta. Adaptado de *J. Chem. Inf. Model. 2006, 46, 1, 401-415.*

Os fatores de enriquecimento variam muito de recetor para recetor. Também dependem do quão semelhantes (estruturalmente) são os ligandos dos não-ligandos. Quando os últimos são razoavelmente semelhantes (o caso mais difícil para efetuar a ordenação) os fatores de enriquecimento variam desde cerca de 3-4 até cerca de 100. A melhor função de pontuação será aquela que obtiver um maior fator de enriquecimento.

A título de exemplo ilustra-se na figura II.8 a deteção de ligandos em função da fração da base de dados considerada. Como se pode ver, quanto maior for a fração da base de dados pesquisada maior é o número de ligandos detetados, mas também maior é o número de falsos positivos (não-ligandos que têm melhor pontuação que ligandos). O exemplo refere-se ao recetor de androgénio humano . O fator de enriquecimento a 1% é de 34 vezes, o que é bastante razoável, em especial num processo puramente automatizado.

2.3. CONCLUSÃO.

O método de ancoragem molecular é uma ferramenta de grande utilidade para a química, bioquímica e biologia molecular. Permite encontrar a pose de associação de um complexo proteína:ligando a partir da estrutura da proteína e do ligando separadamente. Trata-se de um método de utilização não demasiado complexa mas que exige um conhecimento profundo por parte do utilizador. A sua capacidade preditiva é bastante razoável mesmo sem intervenção de um perito mas, para se obterem previsões realmente fiáveis, é necessário escolher cuidadosamente a função de pontuação mais adequada ao sistema em estudo, e conduzir o processo de pesquisa de forma a respeitar todo o conhecimento estrutural e mecanístico pré-existente.

Com o desenvolvimento algorítmico e o progresso tecnológico na área da informática, é possível que o método de ancoragem molecular continue a progredir em exatidão e fiabilidade ao mesmo ritmo, ou a maior ritmo, do que tem vindo a progredir nas últimas décadas.

CAPÍTULO III
TRIAGEM VIRTUAL

3.1. INTRODUÇÃO.

Este capítulo pretende fornecer uma introdução a uma técnica computacional fundamental no processo de descoberta de fármacos- a triagem virtual (em inglês virtual *screening, VS*). Neste capítulo focamo-nos na sua variante mais comum e eficiente, a triagem virtual baseada na estrutura do recetor (em inglês *receptor-based virtual screening*), embora existam outras variantes que são utilizadas quando a estrutura do recetor não é conhecida[5].

A larga maioria dos fármacos são moléculas pequenas (massa molecular relativa abaixo de 500, geralmente menos de 80 átomos) que se associam a recetores biológicos com elevada afinidade. Um recetor biológico não é mais que uma molécula que converte a associação de um ligando numa dada atividade. Os recetores biológicos são, na sua grande maioria, proteínas, dividindo-se essencialmente entre enzimas e proteínas transmembranares[6]. Os recetores são muito maiores que os fármacos que a eles se associam. Os recetores têm muitos milhares de átomos.

Um conceito fundamental na biologia molecular é que a função de um dado recetor é determinada pela sua estrutura. Ou seja, a estrutura define a função. Este conceito esteve na raiz do enorme esforço de determinação estrutural dos recetores macromoleculares, que dura desde há mais de 60 anos, essencialmente através da cristalografia de raios-x. Desvendar a estrutura não seria muito mais do que uma curiosidade académica. Desvendar a função permitiria compreender a fisiologia dos seres vivos, a "maquinaria da vida", a origem da vida, a natureza da vida, e com isso cumprir a mais antiga ambição da ciência: suster a vida, prolongar a vida, e adiar e evitar a morte.

Esse esforço materializou-se em 1952, através da primeira determinação da estrutura de um recetor biológico: a enzima mioglobina do cachalote. Tal empreendimento resultou na atribuição do prémio Nobel da Química ao seu autor em 1962. Desde então foram já depositadas cerca de 180 mil estruturas de proteínas no *Protein Data Bank*, das quais cerca de 80 mil são de proteínas diferentes.

A comunidade científica rapidamente percebeu como o conhecimento da forma de um recetor (e não apenas o conhecimento da sua função) poderia ser usado como ferramenta de intervenção terapêutica, em especial quando a função de um recetor estava desregulada ou tinha um papel na patologia a tratar: criando um ligando exógeno que se associe ao recetor com elevada afinidade, estimulando ou inibindo a sua função. Esta estratégia transformou-se numa ideia central da indústria farmacêutica: descobrir pequenas moléculas[7] que se associassem fortemente a recetores biológicos de interesse e que, através disso, controlassem

[5] Quando a estrutura do recetor é desconhecida pode-se efetuar "triagem virtual baseada no ligando"- procurar moléculas que sejam estruturalmente similares a ligandos conhecidos. No entanto, a triagem baseada na estrutura do recetor é mais eficiente e é preferível.

[6] Alguns fármacos atuam no DNA, RNA, ou lípidos, mas são uma minoria, e não serão aqui discutidos.

[7] Os fármacos não têm necessariamente de ser moléculas pequenas (Mr<500), e nem todos o são. No entanto, as pequenas moléculas são geralmente possíveis de administrar oralmente, que é a forma mais simples e confortável de administração. Quando os fármacos são moléculas grandes (Mr>>500) a sua administração ten geralmente de ser feita intravenosamente, o que é sempre desconfortável e de evitar.

a sua função no sentido desejado. A essas pequenas moléculas chamamos fármacos (**Figura III.1**).

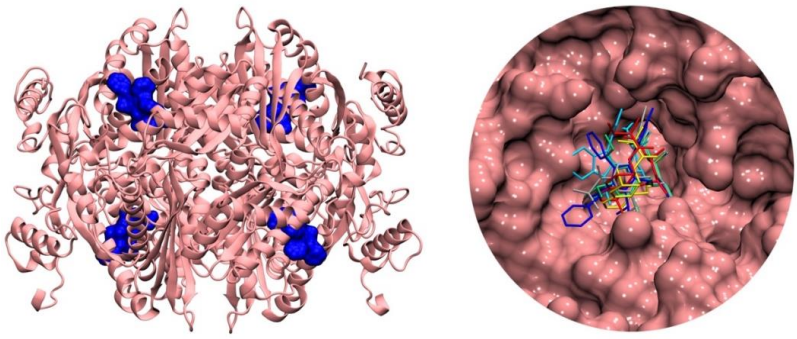


Figura III.1: Os fármacos são geralmente pequenas moléculas que se associam a grandes recetores biológicos. O exemplo ilustra a enzima Redutase do 3-hidroxi-3-metil-glutaril-CoA (HMGR), fundamental para a biossíntese de colesterol. Os fármacos denominados "estatinas" são inibidores competitivos desta enzima, ligando-se ao seu centro ativo com elevada afinidade e bloqueando assim a biossíntese de colesterol. Desta forma, administrando uma dose apropriada de uma estatina, consegue-se descer o a concentração de colesterol de doentes hipercolesterolémicos para níveis adequados. Esquerda: a porção citosólica da enzima HMGR com o fármaco atorvastatina ligado no centro ativo. Direita: um detalhe da mesma enzima com seis fármacos diferentes, todos da família das estatinas, ligadas no centro ativo.

Ao longo das décadas iniciais da determinação estrutural de proteínas, sempre se acreditou que seria possível desenhar novos ligandos de alta afinidade somente a partir da observação da estrutura tridimensional dos seus recetores. No entanto, essa tarefa revelou-se extremamente difícil e minada por uma grande dose de insucesso.

Os métodos experimentais para descoberta de fármacos foram evoluindo, no sentido da automatização e massificação na procura de ligandos. Um novo método emergiu, denominado triagem de alto rendimento (do inglês *high throughput screening, HTS*). A triagem de alto rendimento utiliza robots e vastas bibliotecas de compostos para pipetar automaticamente amostras de milhares de compostos, e as dissolver individualmente em meio tampão apropriado, contendo, entre outras coisas, o recetor biológico para o qual se quer encontrar um ligando. A preparação é feita em placas multipoços (tipicamente com 384 a 1536 poços) e o volume de amostra ronda os 50 nl por poço (**Figura III.2**).

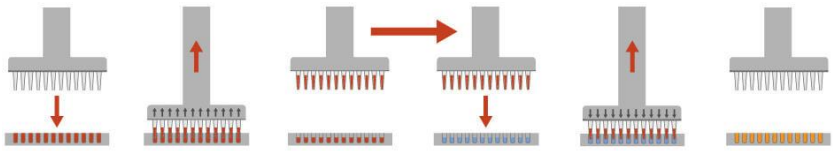


Figura III.2: Um robot de um sistema de HTS retira uma série de amostras de uma biblioteca de compostos e coloca-as numa placa multipoços de grande capacidade (por ex. 1536 poços). A placa é depois conduzida roboticamente a um leitor de placas para medição de atividade. Desta forma simples se podem testar 1536 compostos numa única placa, de forma muito rápida. O processo é repetido centenas ou milhares de vezes, com

amostras diferentes, até triar completamente a biblioteca de compostos, que pode ter vários milhões de compostos. Imagem gentilmente cedida por Singer Instruments, Roadwater, Somerset, TA23 0RE, UK.

Em seguida faz-se uma medição de afinidade recetor-ligando rápida e relativamente eficiente num leitor de placas multipoço, recorrendo tipicamente a métodos espectroscópicos, como por exemplo à espectroscopia de fluorescência. Um sistema moderno de HTS permite triar mais de 100 mil compostos por dia. Um sistema de ultra-HTS moderno consegue triar 300 mil compostos por dia.

Desta forma, consegue-se triar uma biblioteca com 4 milhões de compostos em apenas 2-6 semanas. Esta tecnologia tem, no entanto, duas grandes limitações: o seu elevado custo financeiro e alguma falta de exatidão. De facto, os compostos mais promissores detetados pelo HTS, que são em seguida testados rigorosamente (geralmente menos de 1000, denominados hits) revelam-se frequentemente como falsos positivos. O HTS também é assobrado por um número indesejável de falsos negativos, e não é incomum que os melhores ligandos passem indetetados por uma campanha de HTS. Apesar destas limitações, o HTS é uma técnica de eleição da indústria farmacêutica, uma vez que consegue encontrar ligandos de boa afinidade em aproximadamente 50% dos projetos em que é aplicada.

Paralelamente, o advento e maturação da computação em bioquímica fez nascer a esperança de que o cálculo físico explícito das interações entre ligandos e recetores, permitisse por fim prever a estrutura de ligandos de alta afinidade, e libertasse definitivamente a biologia molecular do tormento da difícil, trabalhosa e pouco frutífera pesquisa e experimentação empírica de ligandos. A ideia de realizar triagens virtuais esteve desde cedo na mente dos bioquímicos computacionais. A triagem virtual segue um esquema idêntico à HTS, com a diferença de que tudo é feito no computador. *Software* de simulação molecular testa, de forma sistemática, grandes números de compostos químicos (vários milhões), no sentido de identificar aqueles que se ligam com alta afinidade a um dado recetor biológico de interesse. A triagem virtual sempre alimentou a esperança de substituir a HTS por uma ferramenta mais rápida, mais precisa e mais barata. No entanto, desde o seu nascimento que a triagem virtual se depara com obstáculos de grande monta, que curiosamente são muito semelhantes aos obstáculos que a HTS enfrenta:

Primeiro: o espaço químico a partir do qual se constroem as bibliotecas de compostos é muito vasto, mas frequentemente desinteressante do ponto de vista biológico. Um grande esforço tem sido feito no sentido de filtrar os compostos químicos do modo a focar as bibliotecas em compostos que realmente têm interesse biológico, ou seja, compostos que não são tóxicos, que são biodisponíveis[8] e que são relevantes para o alvo em investigação. Este obstáculo é comum à triagem virtual e ao HTS.

Segundo: A estrutura dos recetores biológicos é extremamente complexa e muito flexível (ver **Figura III.1**). As proteínas têm forma de novelos, compostas por milhares de átomos com muita liberdade de movimento e rearranjo. Mudam frequentemente de conformação, e é muito difícil prever as adaptações estruturais que cada recetor sofre quando associado a cada ligando da biblioteca.

[8] Biodisponibilidade é a capacidade de um dado composto atingir a circulação sistémica após administração. Costuma ser definido quantitativamente pela razão entre a concentração do composto no plasma quando administrado no local desejado e quando administrado intravenosamente.

Terceiro: o cálculo da energia livre de associação, a propriedade fundamental para se quantificar a afinidade de uma molécula para um dado recetor[9], é muito complexa. É possível determinar, com bastante exatidão, a diferença de energia livre de associação entre duas moléculas similares. Calcular energias livres de associação absolutas, e com rigor, é extremamente difícil, e praticamente impossível quando se tenta fazê-lo para milhões de moléculas diferentes. Este passo virtual corresponde à medição de atividade dos ligandos no leitor de placas na triagem por HTS. Curiosamente, na HTS a medição de atividade é também um problema – é difícil faze-la simultaneamente com velocidade e rigor.

Perante tais dificuldades, a triagem virtual parecia inicialmente estar destinada ao insucesso. Mas de facto, não foi de todo o caso. Fruto da evolução das metodologias e do poder de computação, a triagem virtual atingiu um enorme sucesso, e um sucesso cada vez maior à medida que as suas técnicas vão sendo melhoradas e refinadas. Curiosamente o mesmo se disse no passado da HTS, que foi inclusivamente considerada como estando na origem de um certo declínio na descoberta de novos fármacos que tem ocorrido nas últimas duas décadas. Ou seja, quer a triagem virtual quer a triagem experimental enfrentaram obstáculos muito significativos, e muitas vezes semelhantes. No entanto, ambas conseguiram ultrapassa-los com sucesso e estabeleceram-se como ferramentas centrais e omnipresentes na indústria farmacêutica, na fase inicial da descoberta de novos fármacos, quando se pretende identificar um primeiro composto com alta afinidade para o recetor em estudo. Mais ainda, estabeleceram-se como técnicas complementares, em que o resultado de ambas em conjunto é mais valioso do que o resultado de cada uma separadamente.

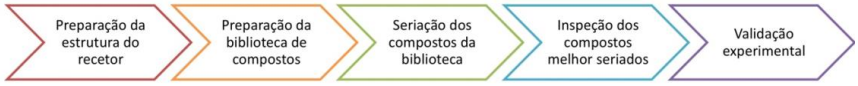
A triagem virtual tende a ser menos automatizável, e consequentemente mais exigente do ponto de vista intelectual, que a HTS, sendo necessários investigadores seniores, com muitos anos de experiência, para a desempenhar com sucesso. A HTS, em contrapartida, é muito mais dispendiosa. O custo da construção de uma infraestrutura de HTS é bastante variável, mas sempre altíssimo. O custo dE HTS subcontratado ronda os 0,1-1 €/poço, dependendo do ensaio, o que representa um custo de 400 mil - 4 milhões de € para uma biblioteca com 4 milhões de compostos. O custo inclui os reagentes, o aluguer de equipamento e o pessoal técnico que executa o HTS. No caso da triagem virtual, o aluguer dos computadores necessários, *software* apropriado e salário dos investigadores custa cerca de 90 mil € para os mesmos 4 milhões de compostos, o que é bastante mais acessível.

A triagem virtual veio democratizar o processo de descoberta de fármacos, que seria um nicho exclusivo de empresas farmacêuticas e de raros grupos de investigação com financiamento principesco caso a HTS fosse o único meio disponível para tal. Abriu as portas da descoberta de fármacos a muitos grupos de investigação sem tanto capital financeiro, mas com grande capital intelectual, e com isso contribuiu imenso para o desenvolvimento de terapias inovadoras para o bem da humanidade.

3.2. ESQUEMA GERAL DA TRIAGEM VIRTUAL.

[9] Uma molécula é considerada como um ligando promissor, de boa afinidade (um composto *hit*) se a sua constante de dissociação (também chamada constante de inibição, K_I) for menor do que 10^{-6} M, ao que corresponde uma energia livre de associação (ΔG_{bind}) abaixo de -8.5 kcal/mol a 37 ºC. Um fármaco, na sua forma final, tem geralmente uma $K_I < 10^{-9}$M e uma $\Delta G_{bind} < -12.8$ kcal/mol.

O **esquema 1** mostra o fluxograma de uma campanha de triagem virtual.



Esquema 1: As várias etapas envolvidas numa triagem virtual. Cada uma será individualmente abordada nos capítulos abaixo.

O primeiro passo é a escolha e preparação do recetor. A estrutura do recetor pode ser obtida por diversas formas, tais como cristalografia de raios-x, ressonância magnética nuclear, ou modelação por homologia, entre outras, mas tem de ser alvo de um tratamento rigoroso posterior.

O segundo passo consiste na preparação da biblioteca de compostos que vai ser triada. Esta deve ser bastante extensa (preferencialmente milhões de compostos), e conter apenas compostos que sejam de interesse biológico, ou seja, que não sejam tóxicos e que sejam biodisponíveis. Existe ainda uma série de requisitos fundamentais para a construção de bibliotecas de compostos que serão estudados mais adiante.

O terceiro passo consiste na escolha do método para prever a afinidade entre o ligando e o recetor em estudo. Este método deverá ser suficientemente rápido para poder triar os milhões de compostos da biblioteca, mas suficientemente preciso para poder discriminar os ligandos dos não ligandos, e efetuar uma seriação dos ligandos em termos de afinidade.

O quarto e último passo consiste na visualização e análise individual dos melhores ligandos, os que serão encaminhados para validação experimental rigorosa. Uma análise humana individual é indispensável, para garantir a maior diversidade química possível e assegurar que interações fundamentais proteína:ligando estão presentes.

A dimensão da lista que é testada experimentalmente com rigor depende da capacidade de teste experimental, mas é usual testar uma ou mais centenas de compostos individualmente.

Vamos agora analisar cada um destes quatro estágios individualmente.

3.2.1. Preparação da estrutura do recetor.

Como é evidente, o sucesso de uma campanha de triagem virtual depende muita qualidade da estrutura do recetor. Estruturas experimentais de ressonância magnética nuclear ou cristalografia de raios-x são, em geral, preferíveis a estruturas modeladas por homologia, a não ser que estas últimas tenham derivado de homólogos com elevada identidade sequencial.

A maior parte das estruturas de recetores disponíveis são de cristalografia de raios-x, pois é esta a melhor técnica para a determinação estrutural de proteínas de média e grande dimensão. É importante garantir que as estruturas cristalográficas estão completas, que não há segmentos por resolver (ou, caso haja, que se localizem em zonas pouco relevantes para a associação com o ligando). É importante ainda garantir que os cofatores, quando existentes, estão presentes na estrutura. A resolução da estrutura é outro fator a ter em conta. Estruturas com má resolução são de evitar em campanhas da triagem virtual. Uma resolução abaixo de 2 Å é ideal.

Um outro aspeto muito importante é a existência de estruturas de recetores com ligandos associados. Ainda que o paradigma de associação molecular dominante seja o de chave-

fechadura (ou seja, o recetor e o ligando mantém a mesma estrutura quando livres e quando associados), um número considerável de pares recetor:ligando obedece ao paradigma do encaixe induzido (ou seja, o recetor e o ligando mudam de estrutura quando se associam). Nestes casos, a estrutura de um recetor sem ligandos não é um bom ponto de partida, pois o seu local de ligação está na geometria apropriada para a forma livre, e não para a forma ligada.

A existência de estruturas do recetor com ligandos associados permite verificar se o local de ligação se reorganiza após a associação. Idealmente, deverão haver estruturas do recetor associadas a vários ligandos diferentes, para se perceber se o local de ligação se adapta, de forma diferente, a cada ligando.

Os casos mais simples são os de recetores que associam os ligandos segundo o paradigma chave-fechadura, pois pode-se assumir que o seu local de ligação permanece na mesma conformação quando se liga a quaisquer moléculas da biblioteca. No caso do encaixe induzido, o processo é um pouco mais complexo. Se estiverem disponíveis estruturas do recetor associado a ligandos, pode-se apagar o ligando do centro ativo e usar essa mesma estrutura do recetor (que está na conformação ativa) para efetuar a triagem virtual. Os casos em que o recetor se reorganiza, de forma diferente, na presença de diferente ligandos, é raro mas bastante problemático. Nesses casos, o problema pode ser resolvido usando um ensamble de conformações do recetor, ou usar um algoritmo de pesquisa conformacional para flexibilizar o recetor, tal como se usa no processo de encaixe molecular de ligandos (**Figura III.3**).

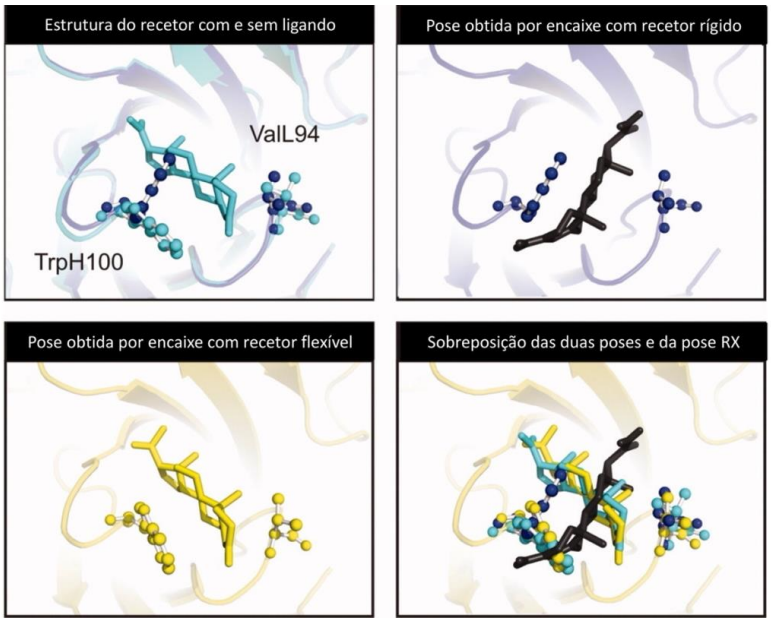


Figura III.3: Um exemplo de encaixe induzido- a progesterona é um ligando de alta afinidade para o anticorpo monoclonal anti-progesterona DB3 (K_i=1·10^{-9} M). O anticorpo for desenvolvido para estudar o papel da progesterona na gravidez; a sua administração provoca uma depleção de progesterona e impede a gravidez em ratos quando administrado após a cópula. Os resíduos Trp100 da cadeia pesada (TrpH100) e Val94 da cadeia leve (ValL94) participam no encaixe induzido da progesterona. O primeiro, em especial, sofre um rearranjo significativo

quando o anticorpo está livre (topo esquerdo, azul ultramarino) e complexado com progesterona (topo esquerdo, azul ciano). O encaixe molecular com recetor rígido, partindo da estrutura cristalográfica que não contém progesterona (1DBA.pdb), gera uma pose incorreta (canto superior direito, progesterona a negro). No entanto, se o encaixe molecular for efetuado considerando um recetor flexível, ou a forma cristalográfica do recetor ligado (1DBB.pdb) a pose obtida já é bastante diferente (canto inferior esquerdo, recetor flexível e progesterona a amarelo). Comparando as duas poses obtidas com a pose cristalográfica (canto inferior direito) pode-se ver que a consideração de um recetor flexível gerou quer uma pose, quer uma estrutura do local de ligação, muito próximas das cristalográficas, ao contrário da pose gerada a partir do recetor não-ligado e sem introduzir flexibilidade no recetor. Esta última estratégia foi incapaz de reproduzir o fenómeno de encaixe induzido. Adaptado com permissão de *Proteins, 2009;74(1):192-206.*

Um problema muito relevante da cristalografia de raios-x é a incapacidade de detetar os átomos de hidrogénio. Tal impede a determinação dos estados de protonação dos resíduos ionizáveis (Arg, Lys, Asp, Glu, His, e por vezes Cys). Como é lógico, a associação receptor:ligando é muito afetada pelo estado de ionização das cadeias laterais dos resíduos próximos, em especial na região de ligação.

É possível calcular, através de algoritmos apropriados (tal como o PropKa ou H++), os estados de ionização dos resíduos de uma proteína. Apesar de os *software* de previsão de estados de ionização funcionarem relativamente bem, são falíveis, e como tal é sempre importante efetuar uma análise visual dos resíduos previstos como tendo estados de ionização atípicos, para tentar garantir que as previsões estão corretas.

O mesmo se passa com os ligandos. Devido à dimensão da biblioteca de compostos, é muito difícil conhecer os estados de ionização de cada composto após a sua associação ao recetor. Os compostos teriam que ser analisados um a um, no local de associação, o que é muito complexo e quase sempre impossível, fazendo com que, por vezes, se utilizem os ligandos em estados de ionização incorretos.

Um outro problema reside na posição/orientação dos grupos amida das cadeias laterais dos resíduos de glutamina e asparagina. Como o oxigénio e o azoto destes grupos são difíceis de distinguir experimentalmente, não é invulgar estarem trocados nas estruturas cristalográficas (**Figura III.4**). Como o carbonilo do grupo amida é um aceitador de pontes de hidrogénio e o grupo NH_2 é um dador de pontes de hidrogénio, a sua troca tem grandes efeitos na associação de ligandos. Como tal, é natural analisar o local de ligação meticulosamente, e procurar a conformação dos grupos amina que maximizam as pontes de hidrogénio estabelecidas. O *software* de encaixe molecular pode fazer este processo de forma automática e relativamente correta.

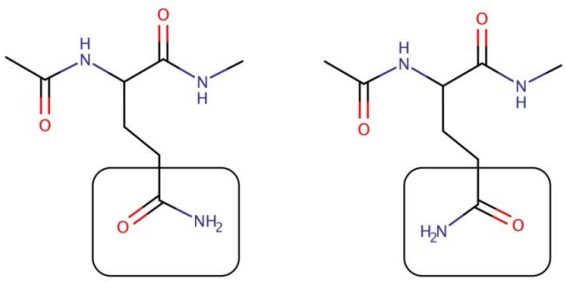


Figura III.4: O grupo amina da cadeia lateral de um resíduo de glutamina mostrado em dois rotâmeros diferentes. Como a cristalografia de raios-x não consegue detetar os átomos de hidrogénio, e os átomos de oxigénio e azoto são difíceis de distinguir, pode ser difícil detetar o rotâmeros correto, não sendo raro encontrar erros na sua atribuição.

Diferentes formas tautoméricas, quando existentes, também devem ser analisadas pormenorizadamente.

A existência de moléculas de água no local de ligação também tem de ser tomada em conta. Muitas vezes as moléculas de água medeiam interações entre ligado e proteína. A não inclusão de moléculas de água pode levar a previsões erradas nas posições e afinidade de ligandos no centro ativo. Existe *software* que permite prever a existência de moléculas de água no centro ativo mas, mais uma vez, a previsão é falível e tem de ser verificada pelo utilizador.

Por fim note-se que o local de ligação deve ser experimentalmente conhecido/determinado. Este poderá ser à superfície da proteína ou enterrado dentro dela e acessível através de um túnel. Deverá ser côncavo, em forma de cavidade, e contendo alguns dadores e aceitadores de pontes hidrogénio e amplas regiões hidrofóbicas. Os locais ligação podem ser centros ativos de enzimas, locais de reconhecimento de recetores transmembranares, locais alostéreos, ou simplesmente cavidades que não tem uma função específica no recetor mas que interferem com a sua função quando um ligando lá se associa. Seja como for, é muito importante ter conhecimento experimental destes locais à partida. Na ausência deste conhecimento, é possível tentar prever a sua localização, mas esta previsão é falível e perigosa, pois se for mal efetuada (o que não é invulgar) irá dirigir a triagem de ligandos para uma cavidade irrelevante para a função do recetor, arruinando todo o processo de triagem virtual desde o seu início.

Apesar de não ser possível garantir que a estrutura do recetor seja a mais correta, a execução criteriosa de todos estes passos, com cuidado e rigor, minimiza os erros estruturais no recetor utilizado e maximiza as hipóteses de sucesso na triagem virtual.

3.2.2. Preparação da biblioteca de compostos.

O sucesso de uma campanha de triagem virtual depende bastante da biblioteca de compostos utilizada. No limite, só é possível obter sucesso se a biblioteca incluir os ligandos que se procura. Se a biblioteca não tiver esses ligandos, o falhanço é garantido, mesmo que todas as outras condições sejam perfeitas.

A biblioteca de compostos é um dos grandes segredos de cada empresa farmacêutica. Qualquer empresa farmacêutica tem a sua própria biblioteca. No meio académico, também muitos grupos tem as suas próprias bibliotecas. Para além destas, existem as bibliotecas públicas que contêm um número enormíssimo de compostos, mais que suficiente para conseguir descobrir e desenvolver um fármaco. Muitas das bibliotecas públicas de compostos têm ainda a vantagem de possuir compostos que podem ser facilmente adquiridos (ou porque existam em armazém ou porque possam ser sintetizados por encomenda). Assim, após a identificação através de triagem virtual, torna-se simples testá-los experimentalmente, uma vez que não é necessário um passo de síntese química.

Existem bibliotecas públicas para todos os gostos. Umas extremamente completas, outras mais focadas numa dada categoria de compostos, outras ainda focadas em medicinas tradicionais ou produtos naturais, ou ainda outras constituídas por fármacos já aprovados e usados na clínica. Seja de que forma for, para garantir o sucesso, uma biblioteca de compostos

deve ter alguns milhões de compostos. Se a biblioteca for reduzida, o sucesso é pouco provável.

As grandes bibliotecas de compostos químicos, tais como o Chemspider, ou PubChem, ou catálogos de fornecedores, como por exemplo o catálogo da Sigma-Aldrich, não são adequadas para triagem virtual. A maior parte das bibliotecas de compostos usadas em triagem virtual são filtradas de acordo com propriedades moleculares potenciadoras de boa farmacocinética. Ou seja, os compostos lá constantes obedecem um conjunto de características físico-químicas que maximizam a probabilidade de os tornarem oralmente biodisponíveis. Estas características podem ser dadas pelas famosas "Regras dos cinco" de Lipinski[10], ou por versões alternativas.

Estas condições não devem ser absolutamente vinculativas, podendo ser permitido que um ligando viole "ligeiramente" as regras de biodisponibilidade. Podem ainda ser relaxadas se se pretende descobrir um fármaco que não seja para administração oral. Adicionalmente, os compostos supostamente tóxicos são retirados das bibliotecas. Os compostos que são conhecidamente problemáticos em ensaio experimentais também costumam ser retirados.

Ou seja, as bibliotecas costumam ter uma lista de compostos químicos diversos, numerosos, não tóxicos, e presumidamente biodisponíveis. As bibliotecas de compostos deste tipo, gratuitas e mais usadas, são a **ChemBL** e **ZINC**. A título de exemplo, vamos explorar a estrutura da segunda, que conta já com 120 milhões de compostos que obedecem às regras de Lipinski (denominados *drug-like* na língua inglesa). A grande maioria destes compostos estão também disponíveis para aquisição. A **Figura III.5** ilustra o painel de escolha de compostos da base de dados ZINC. A **Figura III.6** ilustra a aplicação automática dos filtros mais comuns, o que facilita muito a tarefa do utilizador. Para além destes filtros pré-definidos há inúmeras outras formas de filtrar que o utilizador pode definir com bastante liberdade e flexibilidade.

Para além da filtragem inicial, que remove compostos potencialmente tóxicos, reativos, agregantes, e provavelmente pouco biodisponíveis, um segundo nível de filtragem, específico para o recetor em estudo, pode também ser feito. A esse nível, algumas das questões que podemos colocar a nós próprios são:

1. Se o local de ligação tem um comprimento limitado, deverei manter na biblioteca os compostos muito mais compridos do que o local de ligação?

2. Se a cavidade de ligação tem um volume limitado, deverei manter na biblioteca os compostos muito mais volumosos do que o volume da cavidade?

3. Se o local de ligação tem um cofator metálico (por exemplo, um catião Mg^{2+}) deverei manter na base de dados apenas os ligandos com grupos quelantes que se coordenem fortemente ao cofator?

[10] As regras de Lipinski preveem má absorção ou permeação quando o ligando tem: i) mais do que 5 dadores de pontes de hidrogénio(contados como o n.º de NH e OH); ii) mais que 10 aceitadores de pontes de hidrogénio (contados como o número de N e O); iii) massa molecular relativa maior que 500; iv) O logaritmo da partição 1-octanol /água (logP) é calculado em mais de 5.
Substratos de transportadores biológicos são dispensados de obedecer a estas regras.

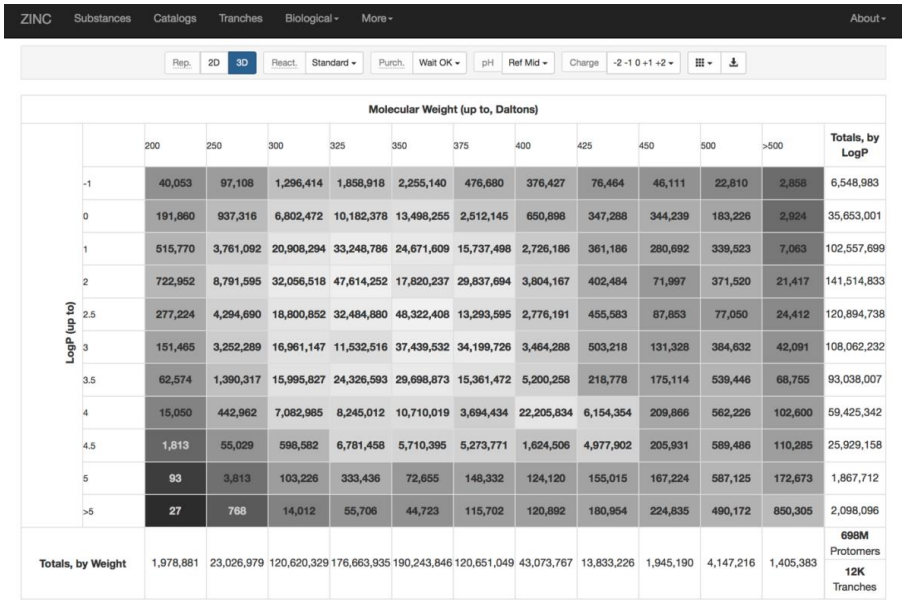
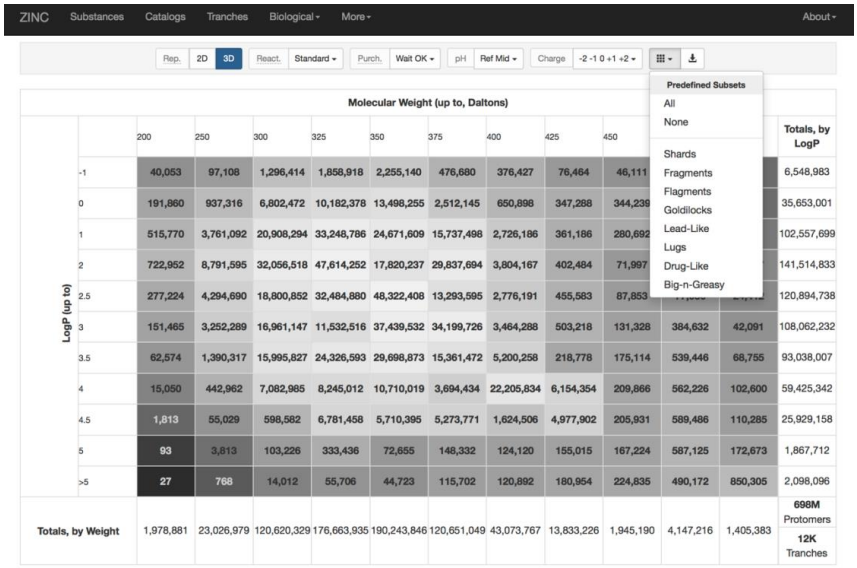


Figura III.5: O painel tranches da base de dados ZINC, que permite dividir os compostos em tranches e descarregar os compostos de uma dada tranche para efetuar triagem virtual. Uma "tranche" é um subconjunto da base de dados em que os compostos obedecem a determinadas características, resultantes dos filtros aplicados. Por exemplo, a figura permite imediatamente descarregar as tranches cujas moléculas estão compreendidas entre determinados limites de massa molecular relativa e de log P. O log P (logaritmo da partição do composto numa mistura bifásica 1-octanol/água) é uma medida da hidrofobicidade da molécula, e como tal é fundamental para a sua associação ao alvo e solubilidade no plasma. Os números dentro de cada célula correspondem à quantidade de compostos que obedecem aos respetivos filtros de logP e massa molecular relativa. Para além destas características, o utilizador pode filtrar, de forma muito simples, os compostos imediatamente disponíveis para aquisição (opção Purch./Wait OK), o pH a que os ligandos estão sujeitos (opção pH) e a carga do ligando (opção -2 -1 o +1 +2). Adaptado com permissão de *Chem. Inf. Model. 2015, 55, 11, 2324-2337.*

4. Se o local de ligação é predominantemente negativo, deverei excluir da base de dados os ligandos que tenham carga negativa a pH fisiológico?

5. Se existirem vários ligandos conhecidos para o meu recetor, e todos tiverem massa molecular e log P semelhante, deverei excluir da base de dados todos os ligandos que tenham massa molecular e log P significativamente diferente dos ligandos conhecidos?

A nossa opinião é que estas filtragens adicionais, e outras, devem preferencialmente ser evitadas. Um ligando pode ser mais volumoso do que o local de ligação, mas "encaixar" uma região na cavidade e deixar a outra região fora da cavidade de ligação, e ainda assim conseguir boa afinidade e ser um bom ponto de partida para o desenvolvimento de fármacos. Pode ser mais comprido que o local de ligação, mas conseguir "acomodar-se" através de um rearranjo conformacional. Um ligando pode ser negativo a pH fisiológico, mas neutro dentro do microambiente do local de ligação.



LogP (up to)	200	250	300	325	350	375	400	425	450	Totals, by LogP
-1	40,053	97,108	1,296,414	1,858,918	2,255,140	476,680	376,427	76,464	46,111	6,548,983
0	191,860	937,316	6,802,472	10,182,378	13,498,255	2,512,145	650,898	347,288	344,239	35,653,001
1	515,770	3,761,092	20,908,294	33,248,786	24,671,609	15,737,498	2,726,186	361,186	280,692	102,557,699
2	722,952	8,791,595	32,056,518	47,614,252	17,820,237	29,837,694	3,804,167	402,484	71,997	141,514,833
2.5	277,224	4,294,690	18,800,852	32,484,880	48,322,408	13,293,595	2,776,191	455,583	87,853	120,894,738
3	151,465	3,252,289	16,961,147	11,532,516	37,439,532	34,199,726	3,464,288	503,218	131,328	384,632	42,091	108,062,232
3.5	62,574	1,390,317	15,995,827	24,326,593	29,698,873	15,361,472	5,200,258	218,778	175,114	539,446	68,755	93,038,007
4	15,050	442,962	7,082,985	8,245,012	10,710,019	3,694,434	22,205,834	6,154,354	209,866	562,226	102,600	59,425,342
4.5	1,813	55,029	598,582	6,781,458	5,710,395	5,273,771	1,624,506	4,977,902	205,931	589,486	110,285	25,929,158
5	93	3,813	103,226	333,436	72,655	148,332	124,120	155,015	167,224	587,125	172,673	1,867,712
>5	27	768	14,012	55,706	44,723	115,702	120,892	180,954	224,835	490,172	850,305	2,098,096
Totals, by Weight	1,978,881	23,026,979	120,620,329	176,663,935	190,243,846	120,651,049	43,073,767	13,833,226	1,945,190	4,147,216	1,405,383	

Figura III.6: O menu predefined subsets do painel tranches da base de dados ZINC (aberto na figura) permite aplicar os conjuntos de filtros mais comuns. Por exemplo, o filtro drug-like escolhe apenas os compostos que obedecem às regras de Lipinski; o filtro lead-like[11] escolhe compostos mais pequenos, e mais facilmente otimizáveis em passos posteriores do desenvolvimento do fármaco, uma vez que se demonstrou que os fármacos aumentam de dimensão e complexidade ao longo do seu aperfeiçoamento posterior; o filtro Big-n-greasy[12] inclui os compostos grandes e muito lipofílicos, ideais para recetores cujo local de ligação tenha essas características e para interfaces proteina:proteina. Adaptado com permissão de *Chem. Inf. Model. 2015, 55, 11, 2324-2337*.

Filtrar compostos no sentido de serem similares a ligandos conhecidos reduz a diversidade das soluções encontradas. Pode-se encontrar um ligando um pouco melhor que os existentes, mas ainda assim bastante semelhante a eles. Por fim, filtrar os compostos que não são quelantes de cofatores metálicos pode eliminar estruturas que se liguem muito bem ao local de ligação e que, posteriormente, podem ser modificadas quimicamente no sentido de se introduzirem os quelantes apropriados.

Ou seja, o ideal, a nosso ver, é ter uma biblioteca de compostos o mais completa possível, e usar um método computacional que consiga prever a afinidade de ligandos o mais rigorosamente possível, deixando que esse método teste todos os compostos e seja o verdadeiro e suserano filtro.

Esse método é o encaixe molecular.

3.2.3. Seriação dos compostos da biblioteca.

A triagem da biblioteca de compostos é feita através de encaixe molecular. Numa primeira fase, usam-se estruturas cristalográficas de ligandos conhecidos associados ao

[11] Os compostos lead-like foram definidos como tendo massa molecular relativa entre 150 e 350, log P calculado menor que 4, e não mais do que 3 dadores e 6 aceitadores de pontes de hidrogénio.

[12] Os compostos big-n-greasy foram definidos como tendo log P calculado entre 2 e 6 e massa molecular relativa entre 300 e 600.

recetor, se disponíveis, para afinar o método de previsão da pose. Apaga-se o ligando e efetua-se o encaixe novamente, para ver se o método de encaixe prevê corretamente a pose. Este procedimento foi já abordado no capítulo sobre encaixe molecular.

Se o método de encaixe se revelar adequado, e reproduzir a pose cristalográfica do ligando, passa-se à segunda fase de afinação: avaliar a capacidade de seriação do método. Para tal, reúne-se um conjunto de ligandos com boa afinidade para o recetor em estudo, e usa-se sofware apropriado para gerar um conjunto de "imitações" (em inglês, decoys)-moléculas estruturalmente parecidas com o ligando mas que elas próprias não são ligandos com boa afinidade . Geram-se 50 imitações para cada ligando. Se estiverem disponíveis 5 ligandos para o recetor de interesse constrói-se assim uma pequena biblioteca de 255 compostos, dos quais 5 são ligandos e 250 são imitações. Em seguida, efetua-se o encaixe molecular das 255 moléculas e faz-se uma seriação por pontuação.

Se a função de pontuação fosse perfeita, os 5 ligandos apareceriam nos 5 primeiros lugares. No entanto, as funções de pontuação são aproximadas, em especial porque precisam de pontuar rapidamente milhões de moléculas, e como tal é improvável que os 5 ligandos ocupem as 5 primeiras posições... mas é expectável que sejam, em geral, melhor pontuados que as imitações!

Como é que vamos quantificar a "qualidade da seriação"? Existem duas métricas, o fator de enriquecimento (do inglês, *enrichment factor, EF*) e a curva ROC (do inglês, *receiver operating characteristic*).

O fator de enriquecimento permite responder à seguinte questão: O que é que se ganha exatamente com uma triagem virtual? Quão melhor é fazer uma triagem virtual em relação a não fazer uma triagem virtual e escolher compostos da biblioteca aleatoriamente?

De uma forma quantitativa, se se fizer uma triagem virtual e se escolher de uma fração (digamos x %) dos compostos melhor pontuados, quão mais provável é encontrar bons ligandos do que retirar compostos da base de dados aleatoriamente?

O fator de enriquecimento a x% ($EF_{x\%}$) é precisamente dado por esta razão. Vamos ver um exemplo concreto, com x=10%. Imaginemos que os ligandos ficam seriados nas posições 2, 5, 8, 14, e 30. Os 10% de compostos melhor pontuados são os 255*10% = 25,5 compostos. A percentagem de ligandos nesse universo de 25,5 compostos é de 4/25,5*100= 15,7 % (o quinto ligando, na posição 30, já não consta dos 10% melhor pontuados). A percentagem de ligandos que se obtém escolhendo aleatoriamente compostos da base de dados é de 5/255*100=2,0%. Com base nisto, o $EF_{10\%}$ é de 15,7/2,0= **7,9**.

Em conclusão, se fizermos uma triagem virtual e retirarmos os 10% de compostos melhor pontuados, vamos encontrar quase 8 vezes mais ligandos do que se não fizermos a triagem e escolhermos aleatoriamente.

Um valor de $EF_{10\%}$=8 é bom? Sim, é razoável. Mas é possível obter melhores resultados e conseguir valores de EF muito superiores. Para bases de dados muito grandes costuma-se calcular o $EF_{1\%}$, e não é invulgar conseguir valores de $EF_{1\%}$=20, ou superiores. Os maiores valores de $EF_{1\%}$ que se conseguem atingir são ≈100.

A curva ROC é a segunda alternativa para medir a qualidade da seriação; faz uma análise mais detalhada e mais completa do que o fator de enriquecimento. Antes de definir o conceito de curva ROC, devemos definir os seguintes conceitos:

- Positivo ou "ligando": molécula que se liga ao recetor; negativo ou "não-ligando": molécula que não se liga ao recetor.
- Verdadeiro positivo (VP): molécula classificada pelo método de análise (neste caso o encaixe molecular) como ligando e que de facto é um ligando.
- Falso positivo (FP): molécula classificada pelo método de análise (neste caso o encaixe molecular) como ligando mas que na verdade é um não-ligando.

A curva ROC obtém-se representando a razão de VP/total de ligandos (rácio de VP) em função de FP/total de não-ligandos (rácio de FP). Idealmente, o método de pontuação deveria detetar apenas os VP e nunca os FP. No entanto, quanto maior a desproporção entre ligandos e não-ligandos (os não-ligandos sendo muito mais numerosos) mais difícil se torna distinguir os ligandos dos não-ligandos, e a deteção de ligandos (VP) é feita cada vez mais à custa de cometer erros (FP).

A curva ROC tem geralmente o aspeto mostrado na **Figura III.7**:

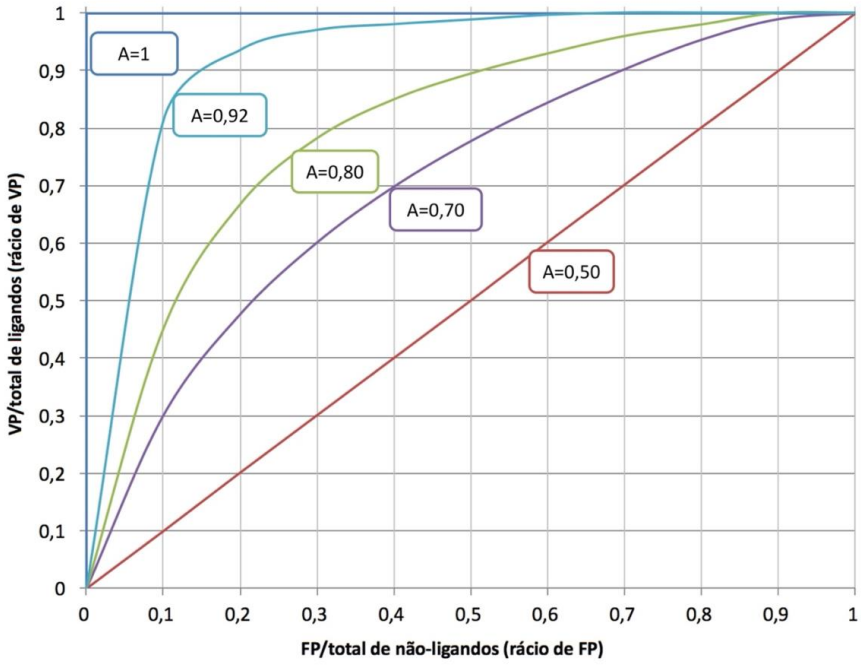


Figura III.6: Uma curva ROC aplicada a um teste de performance de triagem virtual. A área das respetivas curvas está assinalada. A curva vermelha representa o resultado de uma escolha aleatória. O curva azul-escura representa o resultado de um método "perfeito", que não identifica falsos positivos. As curvas roxa, verde e ciano representam casos intermédios, por ordem de precisão de seriação crescente.

Uma seriação perfeita reconhece todos os ligandos (VP) sem classificar erradamente nenhum não-ligando como sendo ligando (FP). Assim, o rácio de VP é 1 e o rácio de FP é zero (linha azul escura). Um método aleatório reconhece, com igual probabilidade relativa, VP e

FP, pelo que resulta em rácios de VP e FP iguais, tal como representado pela linha diagonal vermelha. Um método de seriação é tanto melhor quanto maior o rácio de VP em relação ao rácio de FP (curvas roxa, verde e ciano, por ordem de qualidade crescente). A curva ciano é especialmente favorável porque se refere a uma seriação em que é possível detetar bastantes ligandos sem incorporar na lista muitos não-ligandos. Por exemplo, permite detetar 82% dos ligandos apenas introduzindo 10% de não-ligandos na escolha.

Para quantificar a qualidade da seriação calcula-se a área da curva ROC. Quanto maior a área, melhor é a seriação. No caso ideal, o valor da área é 1 (linha azul-escuro). No caso aleatório, o valor da área é 0,5 (linha vermelha). A área das curvas cresce no sentido roxo<verde<ciano. As áreas estão assinaladas na **Figura III.7**. Em testes de triagem virtual como o que exemplificámos aqui, não é invulgar obter áreas >0,8.

A questão fundamental é: há alguma forma de melhorar a qualidade da seriação, de melhorar os valores do fator de enriquecimento e da área da curva ROC?

Nesta fase preliminar da triagem virtual, é costume tomar duas medidas para maximizar a qualidade da seriação: em primeiro lugar, experimentam-se várias funções de pontuação. Como todas as funções de pontuação são aproximadas, algumas são mais adequadas ao nosso próprio sistema do que outras. Escolhem-se aquelas que resultem em EF e áreas da curva ROC maiores. Em seguida, faz-se uma função de pontuação de consenso: efetua-se uma combinação linear das funções de pontuação com maior EF e área da curva ROC, e por regressão multilinear encontram-se os coeficientes da combinação linear que maximizar EF e a área da curva ROC.

Estando o processo otimizado, quer em termos de previsão da pose, quer em termos da seriação, usa-se o método que conduziu a melhores resultados para seriar a biblioteca de compostos completa, recorrendo a supercomputadores, para fazer o encaixe molecular e pontuação por consenso dos milhões de compostos nela constantes. Este processo é muito exigente, do ponto de vista computacional, não sendo invulgar necessitar de centenas de milhares de horas de computação.

3.2.4. Visualização e inspeção humana dos compostos melhor seriados.

O último passo é a "colheita" manual de compostos. Do passo anterior resultou uma seriação completa da base de dados. Temos, portanto, alguns milhões de compostos seriados por ordem de pontuação. Se a triagem virtual fosse um método exato, isso significaria que teríamos uma seriação da base de dados em função da afinidade para o recetor. No entanto, a estrutura do recetor tem erros e imprecisões, experimentais e de modelação. Os estados de protonação dos ligandos podem ter erros. A pose prevista pelo encaixe molecular tem imprecisões. E por fim, as funções de pontuação têm imprecisões. Se assim não fosse, se os diversos passos do método não tivessem erros, e tivessem imprecisões muito pequenas, poderíamos desenvolver fármacos para todos os recetores conhecidos em muito pouco tempo. Curaríamos quase todas as doenças. A pandemia por Covid-19 poderia ser resolvida de forma muito mais acelerada porque já teríamos um batalhão de fármacos experimentais que funcionariam na perfeição in vitro, em meio tampão, abrindo caminho ao seu funcionamento eficiente in vivo. Mas ainda não se conseguiu chegar a esse estado de desenvolvimento. Ou seja, os resultados da triagem virtual, tais como os resultados experimentais de HTS, são **falíveis**. O composto melhor seriado pode até ser um falso positivo. De facto, é certo que vão

haver muitos falsos positivos na seriação. Se medirmos o sucesso como a percentagem de verdadeiros positivos entre a lista que é testada experimentalmente com rigor (isto é, testada numa grande gama de concentrações e ao longo do tempo, com vários replicados) e não é invulgar que **90% dos compostos escolhidos sejam falsos positivos. Isso significa que só cerca de 10% dos compostos seriados no topo da lista serão efetivamente bons ligandos.**

Estamos perante um fracasso, um falhanço dos métodos de triagem virtual? É claro que não! O que isto significa é que precisamos de testar experimentalmente, com rigor, 100 compostos para poder identificar 10 ligandos de alta afinidade. Se testarmos 500 compostos podemos identificar 50 ligandos de alta afinidade. E quando nos referimos a "alta afinidade", em geral referimo-nos a compostos com uma constante de inibição na escala μM (10^{-6} M) E tal é mais que suficiente para progredir com a linha de descoberta de fármacos.

Se não tivéssemos feito a triagem virtual, e tentássemos testar 500 compostos aleatoriamente, a probabilidade de encontrar ligandos era praticamente zero. Com a triagem virtual, podemos encontrar facilmente 50. Se fizéssemos a triagem experimental, HTS, a percentagem de sucesso, medida desta forma, não seria significativamente diferente.

Por estes motivos, é importante não aceitar cegamente a seriação do método, e fazer uma filtragem final. A filtragem final tem três etapas:

1. Decidir quantos compostos são possíveis de testar experimentalmente com rigor. Tal depende essencialmente dos constrangimentos financeiros e experimentais.
2. Efetuar uma filtragem automática para agrupar os compostos por similaridade química, e escolher um representante de cada grupo. Este passo serve para evitar testar vários compostos muito semelhantes entre si. É preferível identificar ligandos com grande diversidade estrutural e, numa fase mais tardia, efetuar pequenas variações em redor da sua estrutura.
3. Observar os candidatos para teste, um a um, com um *software* de visualização molecular, para perceber as interações que eles fazem com o local de ligação no recetor e, por intuição química, escolher, de entre os melhores, quais efetivamente devem ser testados. Neste passo é usual observar 250-500 compostos para depois escolher metade, ou um terço para teste.

Findo este processo, a triagem virtual está terminada, e o próximo passo é a validação experimental.

3.3. Conclusões.

A descoberta de fármacos aposta cada vez mais na massificação, na automatização e na triagem, quer virtual quer experimental. A triagem virtual tem vindo a assumir uma importância crescente, devido ao seu enorme progresso em termos de eficiência e precisão. O melhoramento dos algoritmos em que se baseia, bem como a evolução constante do *hardware* computacional, traçam uma perspetiva otimista para a triagem virtual – o futuro só poderá ser melhor que o passado, e o progresso na direção da eficiência e precisão é seguro. A triagem virtual, cada vez mais, contribuirá para promover a saúde, e aliviar o sofrimento da humanidade.

CAPÍTULO IV
MÉTODOS HÍBRIDOS QM/MM

4.1. A química computacional.

Cálculos químicos que fazem a previsão de estruturas, energias e outras propriedades de moléculas conhecidas ou desconhecidas são considerados instrumentos importantes em investigação química. Que outros métodos nos podem fornecer a estrutura molecular, calor de formação, momento dipolar, potenciais de ionização, densidades de carga, ordens de ligação, densidades de spin, etc. em apenas uma experiência? A resposta é de que, à parte os métodos teóricos, nenhum outro o consegue fazer.

A objeção habitual aos cálculos é de que os seus resultados podem não ser tão exatos quanto os experimentais, e podem conter artifícios resultantes de modelos teóricos imperfeitos. No entanto, os pontos fortes e fracos dos métodos teóricos mais vulgares são já muito bem conhecidos e, portanto, estimativas realistas das suas exatidões já foram levadas a cabo. Adicionalmente, existem muitos casos em que os resultados dos cálculos podem ser de mais confiança que os experimentais, em especial quando se estudam de pequenas moléculas em fase gasosa. Por exemplo, a medição do calor de formação de um alcano policíclico é uma tarefa que requer uma precisão experimental elevada e uma amostra de grande pureza. Mesmo assim não há garantia de que o valor obtido esteja correto. No entanto, para um número elevado desse tipo de alcanos os métodos mais modernos de mecânica quântica permitem o cálculo de calores de formação que diferem dos seus valores 'verdadeiros' de menos do que 1 $kcal.mol^{-1}$. E, além disso, a partir do mesmo cálculo, sabemos ainda a estrutura molecular do composto! O custo de um desses cálculos são umas horas de tempo de computação e uma meia hora de preparação dos dados.

A química computacional é um campo da ciência no qual *hardware* e *software* são usados para simular um processo químico ou para calcular uma propriedade química. Inicialmente a química computacional referia-se a aspetos aplicados da mecânica quântica. No entanto, presentemente, a química computacional engloba uma larga variedade de sub-áreas que incluem a mecânica quântica (do inglês *quantum mechanics, QM*), mecânica molecular (do inglês *Molecular Mechanics, MM*), métodos híbridos QM/MM (do inglês, *Quantum Mechanics/Molecular Mechanics*), métodos de Monte Carlo, dinâmica browniana, meio eletrostático contínuo, dinâmica de reações, métodos de análise numérica, inteligência artificial, etc. A evolução na área da química computacional é tão rápida que a comunidade pedagógica (e todas as outras comunidades) tem dificuldade em manter-se atualizada. Esta realidade é muito diferente da introdução de outras tecnologias para as quais a comunidade química se adaptou ao longo de muitos anos. Por exemplo, os primeiros instrumentos da espetroscopia de infravermelho foram incorporados no currículo escolar no início dos anos 60, muitos anos depois da indústria ter começado a usar estes aparelhos. O mesmo pode ser dito acerca de outros métodos instrumentais. Sempre que um método era introduzido no currículo, o instrumento tinha geralmente uma vida média de 5-10 anos antes de se tornar necessário um melhoramento do mesmo.

O nosso objetivo, aqui, é saber como podemos utilizar os métodos híbridos QM/MM para obter respostas a problemas de natureza biológica, de modo a que o utilizador possa utilizar mais um instrumento, adicional aos vários instrumentos sintéticos e analíticos que o químico

já tem, para conseguir uma maior compreensão de um evento químico particular ou para prever novos eventos químicos. Estes eventos químicos vão da síntese de moléculas novas até reações moleculares, passando pela dinâmica de moléculas. Assim, começaremos por referir brevemente os métodos de QM e MM para posteriormente descrever os QM/MM.

4.2 Sistemas químicos a estudar.

Há uma enorme variedade de sistemas químicos. Basicamente eles podem agrupar-se em duas classes, com base na sua dimensão: moléculas pequenas (dezenas-centenas de átomos), ou macromoléculas e sistemas biológicos (milhares de átomos ou mais). Quando decidimos estudar um determinado sistema temos de o incluir numa destas duas categorias para saber quais os métodos teóricos que temos disponíveis para levar a cabo esse estudo. No entanto, qualquer uma daquelas moléculas constitui um sistema microscópico e, como tal, o seu movimento é descrito pelas leis da mecânica quântica e os cálculos que se fazem são chamados de cálculos quânticos. O custo e a precisão dos vários tipos de cálculo são obviamente de grande interesse. A escolha do método associado é talvez o segredo de um cálculo bem sucedido!

4.2.1. Moléculas pequenas e sistemas pequenos.

Consideram-se moléculas pequenas e sistemas pequenos aqueles que têm relativamente poucos átomos e são, de um modo geral, estudados com métodos baseados na Mecânica Quântica[13], tais como o representado na Figura IV.1.

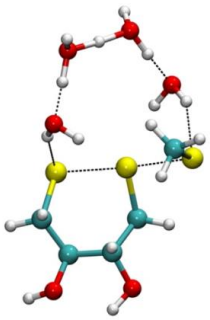


Figura IV.1: Exemplo de um sistema pequeno com menos de 100 átomos. Este sistema é um modelo para estudar uma reação bioquímica muito importante a formação e quebra de ligações dissulfureto em proteínas.

4.2.2. Macromoléculas e sistemas de grande dimensão.

Por outro lado, existem moléculas e sistemas que, por si próprios, são já muito grandes. Exemplos desses sistemas são as proteínas, macromoléculas cujo estudo é muitas vezes apelidado de estudo de moléculas grandes, tais como o representado na Figura IV.2. Os ácidos nucleicos, DNA e RNA, são também exemplos de sistemas muito grandes. Embora haja exceções, de um modo geral, os sistemas biológicos apresentam sempre um número muito

[13] Aos métodos que calculam a energia e propriedades das moléculas com base na mecânica quântica chamamos genericamente **métodos de estrutura eletrónica**.

elevado de átomos sendo, portanto, considerados como grandes. Eles são, de um modo geral, estudados com métodos de Mecânica Molecular[14] e/ou Métodos Híbridos QM/MM.

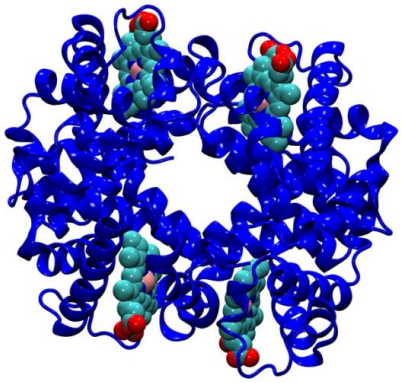


Figura IV.2: Exemplo de um sistema grande com mais de 1000 átomos, a hemoglobina com os cofatores heme ligado no centro ativo.

4.3. QM, mecânica quântica.

In the world of the very small, where particle and wave aspects of reality are equally significant, things do not behave in any way that we can understand from our experience of the everyday world...all pictures are false, and there is no physical analogy we can make to understand what goes on inside atoms. Atoms behave like atoms, nothing else.

— John Gribbin, *In Search of Schrödinger's Cat: Quantum Physics and Reality*

4.3.1. A Equação de Schrödinger.

A Mecânica Quântica é a uma teoria fundamental que descreve a natureza à escala subatómica, ou seja, permite-nos estudar os sistemas microscópicos. Para calcular o valor de propriedades moleculares e compreender os fenómenos subatómicos temos de resolver a equação de Schrödinger[15] (1)

$$\hat{H}\,\Psi = E\,\Psi \qquad (1)$$

na qual $\hat{H}$ é o operador hamiltoniano (ver seção abaixo), Ψ é a função de onda e E é a energia do sistema em questão. A resolução desta equação resulta no conhecimento da energia e da função de onda função do sistema. A função de onda é uma função matemática, cujos argumentos são as coordenadas espaciais dos eletrões e núcleos do sistema atómico, e o valor representa a amplitude de probabilidade de encontrar os núcleos e eletrões numa determinada posição espacial.

[14] A mecânica molecular é um método que utiliza o formalismo da mecânica clássica para calcular a energia e diversas propriedades moleculares através de equações físicas simplificadas, modelando os átomos como sendo esferas unidos por molas.
[15] Aqui referimo-nos a fenómenos independentes do tempo, e sem considerar efeitos relativistas, como é típico e útil na química.

A função de onda é uma função absolutamente extraordinária e única, pois encerra em si toda a informação sobre o sistema. Após conhecer a forma matemática da função de onda, é possível conhecer o valor de todas as propriedades do sistema, com absoluta exatidão, mediante operações matemáticas realizadas posteriormente sobre a função de onda.

Infelizmente, a resolução analítica da equação de Schrödinger de sistemas atómicos só é possível para átomos monoeletrónicos. É pois necessário o desenvolvimento de aproximações que conduzam à solução da equação (1) para um caso geral. A técnica mais comum consistem em

i)trocar a equação de Schrödinger por uma equação parecida, aproximada, mas mais simples de resolver. Em seguida,

ii) resolver a equação aproximada de uma forma também matematicamente aproximada.

Após estas duas aproximações, os resultados não serão mais exatos, mas se as aproximações forem bem feitas os resultados serão ainda bastante próximos dos exatos. Por vezes incorporam-se parâmetros empíricos, experimentais, nas aproximações à equação de Schrödinger, para "compensar" as aproximações feitas. Cálculos quânticos nos quais se realizam as duas aproximações supracitadas, mas que não incorporam parâmetros experimentais são chamados de cálculos *ab initio* (latim para 'desde o início') pois permitem prever tudo sem recurso a experimentação alguma no sistema em estudo[16]. Alternativamente, métodos que combinam teoria com experiência são designados por métodos semi-empíricos e os cálculos associados dizem-se semi-empíricos.

4.3.2. O hamiltoniano molecular.

Uma maneira de chegar à solução da equação (1) para o caso de átomos/moléculas polieletrónicas, é a formulação da hipótese de que os núcleos, numa molécula, estão fixos, em relação aos eletrões[17], considerando apenas o movimento eletrónico. Esta hipótese baseia-se no facto das massas nucleares serem muito maiores do que as eletrónicas.

Esta aproximação, proposta por Born e Oppenheimer, diminui a complexidade da equação de onda completa removendo os termos que envolvem o movimento dos núcleos.

De acordo com a aproximação de Born-Oppenheimer e para uma molécula contendo n núcleos e N eletrões, o hamiltoniano $\hat{H}$, no sistema de unidades atómicas toma a forma da equação 2:

$$\hat{H} = -\frac{1}{2}\sum_{s}^{N}\nabla_s^2 - \sum_{s}^{N}\sum_{A}^{n}\frac{Z_A}{r_{sA}} + \sum_{s>t}^{N}\sum_{t}^{N}\frac{1}{r_{st}} + \sum_{A>B}^{n}\sum_{B}^{n}\frac{Z_A Z_B}{R_{AB}} \qquad (2)$$

na qual se assume que a posição do núcleo A de carga Z_A, é dado pelo vetor $\vec{R}_A$ e que o vetor de posição do eletrão s é $\vec{r}_s$. A quantidade r_{sA}, por exemplo, é a distância entre o eletrão

[16] Os métodos ab initio conseguem calcular as propriedades de um sistema molecular recorrendo apenas a equações da mecânica quântica, e ao valor de propriedades que são independentes do sistema em estudo: o valor da massa e carga dos núcleos e eletrões, o constante de Plank e a permissividade do vazio.

s e o núcleo A, i.e., $r_{sA} = |R_A - r_s|$. Assim, para um sistema com N eletrões, a equação de Schrödinger é uma equação diferencial com respeito a 3N variáveis, só tendo solução analítica se se puder efetuar uma separação de variáveis (isto é, transformar a equação que depende da posição de todos os eletrões do sistema num produto de equações, cada uma dependendo da posição de um único eletrão). No entanto, para sistemas com mais do que um eletrão, não é possível fazer-se a dita separação de variáveis devido à presença do termo eletrónicos $1/r_{st}$ no hamiltoniano $\hat{H}$.

Este termo, denominado o "operador de Coulomb", é o calcanhar de Aquiles de toda a química quântica. É essencialmente ele que nos impede, há cerca de um século, de poder calcular valores exatos para todas as propriedades de qualquer sistema químico. É por causa dele que se empregam as aproximações metodológicas descritas acima, que permitem resolver, de forma aproximada, a equação de Schrödinger aplicada a sistemas com mais do que um eletrão.

4.4. MM, mecânica molecular.

A simulação computacional usando métodos de mecânica molecular tem tido um papel de grande importância nos últimos anos. Este desenvolvimento deve-se aos limites computacionais que, muitas vezes, são insuficientes para a aplicação dos métodos quânticos a sistemas moleculares com um grande número de átomos (mais do que algumas centenas de átomos).

A Mecânica Molecular trata as moléculas como um conjunto de partículas (átomos), com determinada massa, que estão ligadas entre si por um conjunto de molas, e que congregam as propriedades do núcleo e dos eletrões. Para descrever o seu movimento/interações são utilizadas as leis da mecânica clássica. Os átomos, constituem assim a menor unidade do sistema, que passa a ser representado pelas propriedades conjuntas do núcleo e da distribuição média dos eletrões. Desta forma, a energia total do sistema pode ser calculada em função das coordenadas atómicas. A inclusão implícita dos eletrões e de todos os aspetos quânticos do movimento dos núcleos é tratada pela parametrização dos campos de forças.

Um campo de forças tem duas partes fundamentais: i) uma equação que calcula a energia potencial do sistema em função da posição dos seus átomos e ii) conjunto de parâmetros que permite a aplicação dessa equação a sistemas moleculares específicos.

Existem vários campos de forças descritos na literatura, tendo sido cada um deles concebido especificamente para um determinado tipo de moléculas. Eles têm, no entanto, em comum, uma forma geral pela qual a energia total do sistema é decomposta em vários termos independentes, que descrevem a energia necessária à distorção da molécula e as interações entre átomos vizinhos.

$$V_t = V_{lig} + V_\theta + V_\omega + V_{vdw} + V_{el} + V_{cruz} \quad (3)$$

onde, V_{lig} é o termo da energia que corresponde à deformação do comprimento de ligação entre dois átomos, V_θ representa a energia de variação dos ângulos de ligação, V_ω é a

energia de torção para os ângulos diedros, V_{cruz} refere-se a termos cruzados e por último, V_{vdw} e V_{el} descrevem as interações não ligantes de van der Waals e eletrostáticas, respetivamente.

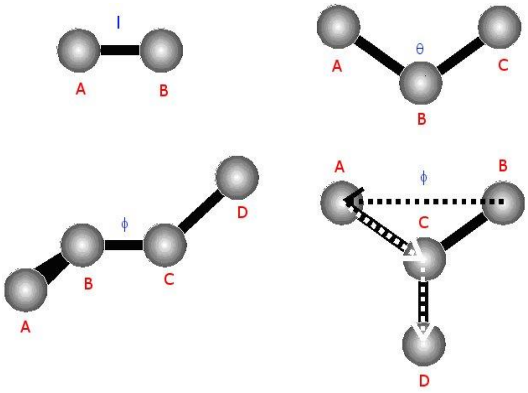


Figura IV.3: l - Comprimento de ligação entre os átomos A e B, θ - ângulo de ligação entre os átomos A, B e C e ϕ - ângulos diedros próprios e impróprios formados pelos átomos A, B, C e D (ângulo entre os planos definidos por A, B e C e por A, C e D).

Os campos de força de mecânica molecular, desenvolvidos especificamente para proteínas, consolidaram-se em exatidão e robustez ao longo das últimas décadas, apresentando presentemente bons resultados. Muito desse desenvolvimento deve-se ao crescente interesse nas simulações de **Dinâmica Molecular** (MD) de sistemas biomoleculares. As simulações de dinâmica molecular geram um conjunto de configurações sucessivas do sistema (trajetória) pela integração das equações de movimento de Newton. Os campos de força são a base fundamental de todas as simulações, tanto de MM como de MD, dado que a sua exatidão e estabilidade dependem da função de energia potencial e dos parâmetros utilizados na equação (3).

Dada a grande dimensão dos sistemas biomoleculares, em simulações de MM e MD, a função de energia potencial usada deve ser simples. Os campos de força mais utilizados em estudos biomoleculares são provavelmente os da família AMBER (Construção Assistida de Modelos e Refinamento de Energia), CHARMM (Química na Mecânica Macromolecular de Harvard) e GROMOS (Sistema de Simulação Molecular de Groningen). A Equação (4) é um exemplo ilustrativo da forma genérica de uma equação de energia potencial usada, com pequenas variações, nos campos de força mencionados acima.

$$E_{Total} = \sum_{Bonds} K_r (r - r_{eq})^2 + \sum_{Angles} K_\theta (\theta - \theta_{eq})^2 + \sum_{Dihedrals} \frac{V_n}{2}\left[1 + \cos(n\phi - \gamma)\right] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right] \quad (4)$$

onde os primeiro e segundo termos contabilizam respetivamente as energias relativas às ligações interatómicas e aos ângulos de ligação; neles, K_r e K_θ referem-se às constantes de força relativas às ligações e aos ângulo. O terceiro termo representa a energia de torção e, o quarto termo, as interações de van der Waals e coulombicas; neles, q_i refere-se às cargas parciais atómicas, ε é a constante dielétrica efetiva, e r_{ij} é a distância entre os átomos *i* e *j*. A maioria dos campos de força usados em estudos moleculares aplica a mesma forma para a função de energia potencial, com termos harmónicos para ligações e ângulos, séries de Fourier para torções, potenciais de Lennard-Jones para as interações de van der Waals e a equação de Coulomb para as interações eletrostáticas, sendo que, as duas últimas são aplicadas entre átomos separados por três ou mais ligações (Equação 4).

4.5. QM/MM, mecânica quântica/mecânica molecular.

O uso de métodos híbridos de QM/MM é uma alternativa muito válida para o estudo de sistemas químicos e biomoleculares, quando os seus tamanhos e complexidade não permitem a aplicação de métodos quânticos puros e os fenómenos que ocorrem no sistema não podem ser descritos corretamente pela mecânica molecular. De facto, apesar do rápido desenvolvimento, em termos de *hardware* e *software*, observado nas últimas décadas, ainda existe uma infinidade de problemas químicos que não podem ser resolvidos por meio da química quântica. Exemplos incluem a quebra e formação de ligações químicas[18], a influência do solvente na reatividade química de certas espécies e o papel catalítico desempenhado pelas enzimas nas reações enzimáticas. Nesses casos, uma descrição explícita do solvente ou da proteína levaria a sistemas de uma dimensão demasiado grande para as metodologias quânticas.

Os sistemas grandes poderiam ser truncados de forma a se construírem modelos pequenos e simples deles. Por exemplo, truncar uma enzima de forma a só deixar o centro ativo e o substrato. Isso permitiria tratar a totalidade do sistema apenas com QM. Obviamente que o uso de um modelo pequeno do sistema real, em QM, não permite que o efeito total da parte restante do sistema seja levado em consideração, o que é uma aproximação de monta. Esta limitação é superada pelos métodos QM/MM.

A filosofia de base por trás das metodologias modernas de QM/MM remonta aos anos 70, em particular ao trabalho pioneiro de Warshel & Levitt, que abriu o caminho para os desenvolvimentos metodológicos que viriam nas décadas seguintes e lhes valeu o Prémio Nobel de 2013. De um modo geral, os cálculos de QM/MM admitem que sistemas químicos de grandes dimensões, incluindo os de natureza biomolecular, podem ser divididos numa região eletronicamente importante que requer um tratamento químico quântico (a região QM) e uma região circundante que admite um tratamento clássico, usando mecânica molecular (a região MM) (**Figura IV.4**).

[18] Uma das grandes limitações da MM em química é de não permitir descrever a quebra e formação de ligações químicas, por não descrever explicitamente os eletrões.

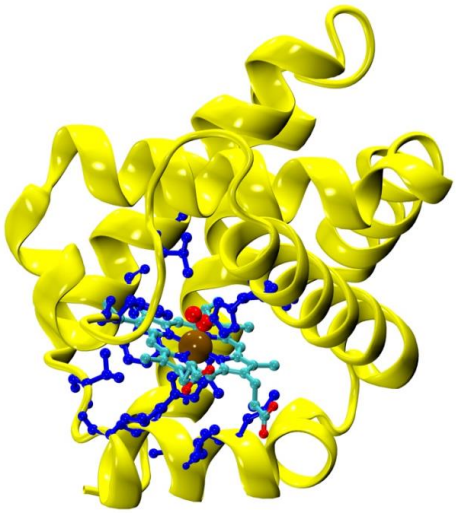


Figura IV.4: Representação das regiões QM e MM num esquema QM/MM. Neste caso trata-se da hemoglobina humana. A região QM está representada atomisticamente, com o centro ativo em azul escuro, o grupo heme e oxigénio coloridos por elemento químico. O cofator metálico (ferro) é representado por uma esfera de maior dimensão e colorido em ocre. A região MM está representada no estilo cartoon, para evidenciar a estrutura secundária da proteína, e colorida a amarelo.

A região QM compreende a parte do sistema diretamente envolvida numa dada reação química em estudo, enquanto a porção MM abrange o restante do sistema, onde eventos como a formação e quebra de ligações não ocorrem. A região MM atua apenas indiretamente na região eletronicamente importante. Assim, essa interação pode ser considerada uma mera perturbação, assumindo duas formas principais: i) estereoquímica, se a região MM impuser a adoção de uma determinada geometria pela porção QM, e ii) eletrostática e de polarização. Assim, de acordo com o princípio subjacente associado aos métodos híbridos QM/MM, uma reação química é tratada como uma transformação envolvendo apenas a região QM, mas que é influenciada pelo ambiente circundante (a região MM), tipicamente o solvente ou a porção restante da proteína no caso de uma reação enzimática.

A literatura na área das técnicas QM/MM é vasta, existindo várias implementações de QM/MM na literatura. As diferenças entre abordagens surgem normalmente do nível da teoria da mecânica quântica empregada e do campo de força da mecânica molecular.

Os métodos mais populares para caracterização da porção de QM em cálculos combinados de QM/MM de sistemas biomoleculares têm sido os métodos semi-empíricos, os métodos de Hartree-Fock e o método da teoria do funcional da densidade (DFT).

4.6. Conclusões.

Os métodos QM/MM permitem aplicar diferentes métodos teóricos a diferentes regiões do sistema. Permitem usar o melhor que cada método teórico tem: a rapidez e capacidade de descrever grandes sistemas da mecânica molecular e a exatidão, versatilidade e capacidade de descrever reações químicas da mecânica quântica. Juntos permitem resolver problemas

que nenhum deles separadamente consegue resolver. Neste caso, como em tantos outros, a união faz a força!

References

1. Berlin, A. & Balakshin, M. *Industrial Lignins: Analysis, Properties, and Applications*. *Bioenergy Research: Advances and Applications* (Elsevier, 2014). doi:10.1016/B978-0-444-59561-4.00018-8.
2. Wrigley, E. A., A, P. T. R. S. & Wrigley, E. A. Energy and the English Industrial Revolution Energy and the English Industrial Revolution. (2013).
3. Brown, M. Making sense of modernity's maladies: health and disease in the Industrial Revolution. *Endeavour* **30**, 108–112 (2006).
4. Morton, D. L. Reviewing the history of electric power and electrification. *Endeavour* **26**, 60–63 (2002).
5. Clark, J. H., Luque, R. & Matharu, A. S. Green Chemistry, Biofuels, and Biorefinery. *Annu. Rev. Chem. Biomol. Eng.* **3**, 183–207 (2012).
6. United Nations. *Kyoto Protocol Reference Manual*. *United Nations Framework Convention on Climate Change* (2008).
7. French, D. Kyoto Protocol to the United Nations Framework Convention on Climate Change. *J. Environ. Law* (1998) doi:10.1093/jel/10.1.215.
8. Lewis, S. L. The Paris Agreement has solved a troubling problem. *Nature* **532**, 283–283 (2016).
9. Damanik, N., Ong, H. C., Tong, C. W., Mahlia, T. M. I. & Silitonga, A. S. A review on the engine performance and exhaust emission characteristics of diesel engines fueled with biodiesel blends. *Environ. Sci. Pollut. Res.* **25**, 15307–15325 (2018).
10. Bryant, L. The Development of the Diesel Engine. *Technol. Cult.* **17**, 432 (1976).
11. Knothe, G. Biodiesel and renewable diesel: A comparison. *Prog. Energy Combust. Sci.* **36**, 364–373 (2010).
12. Yusuf, N. N. A. N., Kamarudin, S. K. & Yaakub, Z. Overview on the current trends in biodiesel production. *Energy Convers. Manag.* (2011) doi:10.1016/j.enconman.2010.12.004.
13. Refaat, A. A. Biofuels from waste materials. in *Comprehensive Renewable Energy* (2012). doi:10.1016/B978-0-08-087872-0.00518-7.

14. Hu, Q. *et al.* Microalgal triacylglycerols as feedstocks for biofuel production: Perspectives and advances. *Plant Journal* (2008) doi:10.1111/j.1365-313X.2008.03492.x.
15. Poppe, J. K., Fernandez-Lafuente, R., Rodrigues, R. C. & Ayub, M. A. Z. Enzymatic reactors for biodiesel synthesis: Present status and future prospects. *Biotechnology Advances* (2015) doi:10.1016/j.biotechadv.2015.01.011.
16. Bušić, A. *et al.* Recent Trends in Biodiesel and Biogas Production. **56**,.
17. Kalscheuer, R., Stölting, T. & Steinbüchel, A. Microdiesel: Escherichia coli engineered for fuel production. *Microbiology* **152**, 2529–2536 (2006).
18. Eriksen, D. T., HamediRad, M., Yuan, Y. & Zhao, H. Orthogonal Fatty Acid Biosynthetic Pathway Improves Fatty Acid Ethyl Ester Production in Saccharomyces cerevisiae. *ACS Synth. Biol.* **4**, 808–814 (2015).
19. Alvira, P., Tomás-Pejó, E., Ballesteros, M. & Negro, M. J. Pretreatment technologies for an efficient bioethanol production process based on enzymatic hydrolysis: A review. *Bioresour. Technol.* (2010) doi:10.1016/j.biortech.2009.11.093.
20. Marriott, P. E., Gómez, L. D. & Mcqueen-Mason, S. J. Unlocking the potential of lignocellulosic biomass through plant science. *New Phytol.* (2016) doi:10.1111/nph.13684.
21. Sun, S., Sun, S., Cao, X. & Sun, R. The role of pretreatment in improving the enzymatic hydrolysis of lignocellulosic materials. *Bioresource Technology* (2016) doi:10.1016/j.biortech.2015.08.061.
22. Rosevear, A. & Kennedy, J. F. Bioconversion of cellulosic substances into energy: Chemicals and microbial proteins. *Enzyme Microb. Technol.* (1979) doi:10.1016/0141-0229(79)90037-1.
23. Sarkar, N., Ghosh, S. K., Bannerjee, S. & Aikat, K. Bioethanol production from agricultural wastes: An overview. *Renewable Energy* (2012) doi:10.1016/j.renene.2011.06.045.
24. Chettri, D., Verma, A. K., Sarkar, L. & Verma, A. K. Role of extremophiles and their extremozymes in biorefinery process of lignocellulose degradation. *Extremophiles* **25**, 203–219 (2021).
25. Cooper, G. M. *The Central Role of Enzymes as Biological Catalyst*. *The Cell: A*

Molecular Approach. (Sunderland, Sinauer Associates, 2000).

26. Lindskog, S. Structure and mechanism of Carbonic Anhydrase. *Pharmacol. Ther.* **74**, 1–20 (1997).
27. Gilmour, K. M. *Transport and exchange of respiratory gases in the blood | Carbonic Anhydrase in Gas Transport and Exchange*. *Encyclopedia of Fish Physiology* vol. 2 (Elsevier Inc., 2011).
28. ARMSTRONG, E. F. Enzymes: A Discovery and its Consequences. *Nature* **131**, 535–537 (1933).
29. Kühne, W. *Verhandlungen Des Naturhistorisch-Medicinischen Vereins Zu Heidelb*. (1877).
30. McGovern, P. E. *et al.* Fermented beverages of pre- and proto-historic China. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 17593–17598 (2004).
31. Dekker, P. Dairy Enzymes. in *Industrial Enzyme Applications* 143–166 (John Wiley & Sons, Ltd, 2019). doi:10.1002/9783527813780.ch2_3.
32. Deák, T. YEASTS. in *Encyclopedia of Food Sciences and Nutrition* 6233–6239 (Elsevier, 2003). doi:10.1016/B0-12-227055-X/01302-X.
33. Araújo, R., Casal, M. & Cavaco-Paulo, A. Application of enzymes for textile fibres processing. *Biocatal. Biotransformation* **26**, 332–349 (2008).
34. Walsh, G. A., Power, R. F. & Headon, D. R. Enzymes in the animal-feed industry. *Trends in Biotechnology* (1993) doi:10.1016/0167-7799(93)90006-U.
35. Cowieson, A. J., Hruby, M. & Pierson, E. E. M. Evolving enzyme technology: impact on commercial poultry nutrition. *Nutr. Res. Rev.* (2006) doi:10.1079/nrr2006121.
36. Cowieson. A, Masey-O´Neill, H. & Bedford, M. Enzymes Beyond Phytase in Poultry Nutrition. *Poult. Res. Found.* (2013).
37. Singh, G., Capalash, N., Kaur, K., Puri, S. & Sharma, P. Enzymes: Applications in Pulp and Paper Industry. in *Agro-Industrial Wastes as Feedstock for Enzyme Production: Apply and Exploit the Emerging and Valuable Use Options of Waste Biomass* (2016). doi:10.1016/B978-0-12-802392-1.00007-1.
38. Srinivasan, M. C. & Rele, M. V. Microbial xylanases for paper industry. *Curr. Sci.* (1999).
39. Bajpai, P. Application of enzymes in the pulp and paper industry. *Biotechnol.*

Prog. (1999) doi:10.1021/bp990013k.

40. d'Errico, C. *et al.* Improved biomass degradation using fungal glucuronoyl—esterases—hydrolysis of natural corn fiber substrate. *J. Biotechnol.* **219**, 117–123 (2016).
41. Abin-Fuentes, A., Mohamed, M. E. S. E.-S., Wang, D. I. C. & Prather, K. L. J. Exploring the mechanism of biocatalyst inhibition in microbial desulfurization. *Appl. Environ. Microbiol.* **79**, 7807–7817 (2013).
42. Shrivastava, A., Shrivastava, N. & Singh, P. K. Enzymes in pharmaceutical industry. in *Enzymes in Food Biotechnology: Production, Applications, and Future Prospects* (2018). doi:10.1016/B978-0-12-813280-7.00034-7.
43. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).
44. International Union of Biochemistry and Molecular Biology, N. C. *Enzyme Nomenclature*. (Academic Press, 1992).
45. Dewan, S. S. *Global Markets for Enzymes in Industrial Applications.* (2018).
46. Khanal, S. K. *Anaerobic Biotechnology for Bioenergy Production: Principles and Applications*. *Anaerobic Biotechnology for Bioenergy Production: Principles and Applications* (2009). doi:10.1002/9780813804545.
47. White, S. W., Zheng, J., Zhang, Y.-M. & Rock, C. O. The Structural Biology of Type II Fatty Acid Biosynthesis. *Annu. Rev. Biochem.* **74**, 791–831 (2005).
48. Maier, T., Leibundgut, M. & Ban, N. The Crystal Structure of a Mammalian Fatty Acid Synthase. *Science (80-.).* **321**, 1315–1322 (2008).
49. Smith, S., Witkowski, A. & Joshi, A. K. Structural and functional organization of the animal fatty acid synthase. *Prog. Lipid Res.* **42**, 289–317 (2003).
50. Grininger, M. Perspectives on the evolution, assembly and conformational dynamics of fatty acid synthase type I (FAS I) systems. *Curr. Opin. Struct. Biol.* **25**, 49–56 (2014).
51. Leibundgut, M., Jenni, S., Frick, C. & Ban, N. Structural basis for substrate delivery by acyl carrier protein in the yeast fatty acid synthase. *Science* **316**, 288–90 (2007).
52. Lomakin, I. B., Xiong, Y. & Steitz, T. A. The Crystal Structure of Yeast Fatty Acid Synthase, a Cellular Machine with Eight Active Sites Working Together. *Cell*

129, 319–332 (2007).

53. Chakravarty, B., Gu, Z., Chirala, S. S., Wakil, S. J. & Quiocho, F. A. Human fatty acid synthase: Structure and substrate selectivity of the thioesterase domain. *Proc. Natl. Acad. Sci.* **101**, 15567–15572 (2004).
54. Paiva, P. *et al.* Animal Fatty Acid Synthase: A Chemical Nanofactory. *Chem. Rev.* **121**, 9502–9553 (2021).
55. Viegas, M. F., Neves, R. P. P., Ramos, M. J. & Fernandes, P. A. Modeling of Human Fatty Acid Synthase and in Silico Docking of Acyl Carrier Protein Domain and Its Partner Catalytic Domains. *J. Phys. Chem. B* **122**, 77–85 (2018).
56. Jenni, S., Leibundgut, M., Maier, T. & Ban, N. Architecture of a fungal fatty acid synthase at 5 A resolution. *Science* **311**, 1263–1267 (2006).
57. Aprahamian, S. A., Arslanian, M. J. & Wakil, S. J. Comparative studies on the kinetic parameters and product analyses of chicken and rat liver and yeast fatty acid synthetase. *Comp. Biochem. Physiol. Part B Comp. Biochem.* **71**, 577–582 (1982).
58. Schweizer, E. & Hofmann, J. Microbial Type I Fatty Acid Synthases (FAS): Major Players in a Network of Cellular FAS Systems. *Am. Soc. Microbiol.* **68**, 501–17, table of contents (2004).
59. Rock, C. O. & Jackowski, S. Forty years of bacterial fatty acid synthesis. *Biochem. Biophys. Res. Commun.* **292**, 1155–66 (2002).
60. Kauppinen, S., Siggaard-Andersen, M. & von Wettstein-Knowles, P. β-ketoacyl-ACP synthase I of Escherichia coli: Nucleotide sequence of thefabB gene and identification of the cerulenin binding residue. *Carlsberg Res. Commun.* **53**, 357–370 (1988).
61. Heath, R. J. & Rock, C. O. The Claisen condensation in biology. *Nat. Prod. Rep.* **19**, 581–596 (2002).
62. D'Agnolo, G., Rosenfeld, I. S. & Vagelos, P. R. Multiple forms of β ketoacyl acyl carrier protein synthetase in Escherichia coli. *J. Biol. Chem.* **250**, 5289–5294 (1975).
63. Garwin, J. L., Klages, A. L., Cronan Jr, J. E. & Cronan, J. E. Structural, Enzymatic, and Genetic Studies of Beta-Ketoacyl-Acyl Carrier Protein Synthases I and Ii of Escherichia Coli. *J. Biol. Chem* **255**, 11949–56 (1980).

64. Tsay, J. T., Oh, W., Larson, T. J., Jackowski, S. & Rock, C. O. Isolation and Characterization of the Beta-Ketoacyl-Acyl Carrier Protein Synthase Iii Gene (Fabh) from Escherichia Coli K-12. *J. Biol. Chem* **267**, 6807–14 (1992).
65. Joshi, A. K., Witkowski, A. & Smith, S. The malonyl/acetyltransferase and β-ketoacyl synthase domains of the animal fatty acid synthase can cooperate with the acyl carrier protein domain of either subunit. *Biochemistry* (1998) doi:10.1021/bi971886v.
66. Zhu, Z. *et al.* Expanding the product portfolio of fungal type I fatty acid synthases. *Nat. Chem. Biol.* **13**, 360–362 (2017).
67. Xu, P., Qiao, K., Ahn, W. S. & Stephanopoulos, G. Engineering Yarrowia lipolytica as a platform for synthesis of drop-in transportation fuels and oleochemicals. *Proc. Natl. Acad. Sci.* **113**, 10848–10853 (2016).
68. Zheng, Y. *et al.* Boosting the free fatty acid synthesis of Escherichia coli by expression of a cytosolic Acinetobacter baylyi thioesterase. *Biotechnol. Biofuels* **5**, 76 (2012).
69. Junne, S. & Kabisch, J. Fueling the future with biomass: Processes and pathways for a sustainable supply of hydrocarbon fuels and biogas. *Eng. Life Sci.* **17**, 14–26 (2017).
70. Kang, M.-K. & Nielsen, J. Biobased production of alkanes and alkenes through metabolic engineering of microorganisms. *J. Ind. Microbiol. Biotechnol.* **44**, 613–622 (2017).
71. Cheon, S., Kim, H. M., Gustavsson, M. & Lee, S. Y. Recent trends in metabolic engineering of microorganisms for the production of advanced biofuels. *Curr. Opin. Chem. Biol.* **35**, 10–21 (2016).
72. Zargar, A. *et al.* Leveraging microbial biosynthetic pathways for the generation of 'drop-in' biofuels. *Curr. Opin. Biotechnol.* **45**, 156–163 (2017).
73. Kerschgens, B., Cai, L., Pitsch, H., Heuser, B. & Pischinger, S. Di-n-buthylether, n-octanol, and n-octane as fuel candidates for diesel engine combustion. *Combust. Flame* **163**, 66–78 (2016).
74. Kumar, B. R. *et al.* A comparative analysis on combustion and emissions of some next generation higher-alcohol/diesel blends in a direct-injection diesel engine. *Energy Convers. Manag.* **119**, 246–256 (2016).

75. Cai, L. *et al.* An experimental and modeling study of n-octanol combustion. *Proc. Combust. Inst.* **35**, 419–427 (2015).
76. Kremer, F., Blank, L. M., Jones, P. R. & Akhtar, M. K. A Comparison of the Microbial Production and Combustion Characteristics of Three Alcohol Biofuels: Ethanol, 1-Butanol, and 1-Octanol. *Frontiers in Bioengineering and Biotechnology* vol. 3 112 (2015).
77. Henritzi, S., Fischer, M., Grininger, M., Oreb, M. & Boles, E. An engineered fatty acid synthase combined with a carboxylic acid reductase enables de novo production of 1-octanol in Saccharomyces cerevisiae. *Biotechnol. Biofuels* **11**, 1–12 (2018).
78. Gabani, P. & Singh, O. V. Radiation-resistant extremophiles and their potential in biotechnology and therapeutics. *Appl. Microbiol. Biotechnol.* **97**, 993–1004 (2013).
79. Adrio, J. L. & Demain, A. L. Microbial enzymes: tools for biotechnological processes. *Biomolecules* **4**, 117–139 (2014).
80. Yu, H. & Huang, H. Engineering proteins for thermostability through rigidifying flexible sites. *Biotechnol. Adv.* **32**, 308–315 (2014).
81. Takano, K. *et al.* Proline effect on the thermostability and slow unfolding of a hyperthermophilic protein. *J. Biochem.* (2009) doi:10.1093/jb/mvn144.
82. Barlow, D. J. & Thornton, J. M. Ion-pairs in proteins. *J. Mol. Biol.* (1983) doi:10.1016/S0022-2836(83)80079-5.
83. Bhuiya, M. W., Suryadi, J., Zhou, Z. & Brown, B. A. Structure of the Aeropyrum pernix L7Ae multifunctional protein and insight into its extreme thermostability. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **69**, 979–988 (2013).
84. Bezsudnova, E. Y. *et al.* Intramolecular hydrogen bonding in the polyextremophilic short-chain dehydrogenase from the archaeon Thermococcus sibiricus and its close structural homologs. *Biochimie* **118**, 82–89 (2015).
85. Gromiha, M. M., Pathak, M. C., Saraboji, K., Ortlund, E. A. & Gaucher, E. A. Hydrophobic environment is a key factor for the stability of thermophilic proteins. *Proteins Struct. Funct. Bioinforma.* (2013) doi:10.1002/prot.24232.
86. Vijayabaskar, M. S. & Vishveshwara, S. Comparative analysis of thermophilic and mesophilic proteins using Protein energy networks. *BMC Bioinformatics*

(2010) doi:10.1186/1471-2105-11-S1-S49.

87. de Oliveira, T. B., Gostinčar, C., Gunde-Cimerman, N. & Rodrigues, A. Genome mining for peptidases in heat-tolerant and mesophilic fungi and putative adaptations for thermostability. *BMC Genomics* (2018) doi:10.1186/s12864-018-4549-5.
88. Haney, P. J., Stees, M. & Konisky, J. Analysis of thermal stabilizing interactions in mesophilic and thermophilic adenylate kinases from the Genus methanococcus. *J. Biol. Chem.* (1999) doi:10.1074/jbc.274.40.28453.
89. Tompa, D. R., Gromiha, M. M. & Saraboji, K. Contribution of main chain and side chain atoms and their locations to the stability of thermophilic proteins. *J. Mol. Graph. Model.* (2016) doi:10.1016/j.jmgm.2016.01.001.
90. Brandl, C. J. & Deber, C. M. Hypothesis about the function of membrane-buried proline residues in transport proteins. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 917–921 (1986).
91. Sansom, M. S. P. Proline residues in transmembrane helices of channel and transport proteins: A molecular modelling study. *Protein Eng. Des. Sel.* (1992) doi:10.1093/protein/5.1.53.
92. Li, S. C., Goto, N. K., Williams, K. A. & Deber, C. M. α-Helical, but not β-sheet, propensity of proline is determined by peptide environment. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 6676–6681 (1996).
93. Mallick, P., Boutz, D. R., Eisenberg, D. & Yeates, T. O. Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds. *Proc. Natl. Acad. Sci.* **99**, 9679 LP – 9684 (2002).
94. Beeby, M. *et al.* The Genomics of Disulfide Bonding and Protein Stabilization in Thermophiles. *PLOS Biol.* **3**, e309 (2005).
95. Ladenstein, R. & Antranikian, G. Proteins from hyperthermophiles: stability and enzymatic catalysis close to the boiling point of water. *Adv. Biochem. Eng. Biotechnol.* **61**, 37–85 (1998).
96. Ladenstein, R. & Ren, B. Reconsideration of an early dogma, saying ‘there is no evidence for disulfide bonds in proteins from archaea’. *Extremophiles* (2008) doi:10.1007/s00792-007-0076-z.
97. Jorda, J. & Yeates, T. O. Widespread disulfide bonding in proteins from

thermophilic archaea. *Archaea* (2011) doi:10.1155/2011/409156.

98. Gírio, F. M. *et al.* Hemicelluloses for fuel ethanol: A review. *Bioresource Technology* (2010) doi:10.1016/j.biortech.2010.01.088.
99. Brumm, P. J., De Maayer, P., Mead, D. A. & Cowan, D. A. Genomic analysis of six new Geobacillus strains reveals highly conserved carbohydrate degradation architectures and strategies. *Front. Microbiol.* (2015) doi:10.3389/fmicb.2015.00430.
100. Collins, T., Gerday, C. & Feller, G. Xylanases, xylanase families and extremophilic xylanases. *FEMS Microbiology Reviews* (2005) doi:10.1016/j.femsre.2004.06.005.
101. Ayoub, A., Venditti, R. A., Pawlak, J. J., Sadeghifar, H. & Salam, A. Development of an acetylation reaction of switchgrass hemicellulose in ionic liquid without catalyst. *Ind. Crops Prod.* (2013) doi:10.1016/j.indcrop.2012.10.036.
102. Balakshin, M., Capanema, E., Gracz, H., Chang, H. min & Jameel, H. Quantification of lignin-carbohydrate linkages with high-resolution NMR spectroscopy. *Planta* **233**, 1097–1110 (2011).
103. Kwiatkowski, J. R., McAloon, A. J., Taylor, F. & Johnston, D. B. Modeling the process and costs of fuel ethanol production by the corn dry-grind process. *Ind. Crops Prod.* **23**, 288–296 (2006).
104. Humbird, D. *et al.* Process Design and Economics for Biochemical Conversion of Lignocellulosic Biomass to Ethanol: Dilute-Acid Pretreatment and Enzymatic Hydrolysis of Corn Stover. *Natl. Renew. Energy Lab.* 1–147 (2011) doi:10.2172/1107470.
105. Arnling Bååth, J., Giummarella, N., Klaubauf, S., Lawoko, M. & Olsson, L. A glucuronoyl esterase from Acremonium alcalophilum cleaves native lignin-carbohydrate ester bonds. *FEBS Lett.* **590**, 2611–2618 (2016).
106. Monrad, R. N., Eklöf, J., Krogh, K. B. R. M. & Biely, P. Glucuronoyl esterases: diversity, properties and biotechnological potential. A review. *Crit. Rev. Biotechnol.* **38**, 1121–1136 (2018).
107. Špániková, S. & Biely, P. Glucuronoyl esterase - Novel carbohydrate esterase produced by Schizophyllum commune. *FEBS Lett.* **580**, 4597–4601 (2006).

108. Mazurkewich, S. *et al.* Structural and biochemical studies of the glucuronoyl esterase otce15a illuminate its interaction with lignocellulosic components. *J. Biol. Chem.* **1**, 1–13 (2019).
109. Arnling Bååth, J. *et al.* Biochemical and structural features of diverse bacterial glucuronoyl esterases facilitating recalcitrant biomass conversion. *Biotechnol. Biofuels* **11**, 1–14 (2018).
110. Pokkuluri, P. R. *et al.* Structure of the catalytic domain of glucuronoyl esterase Cip2 from Hypocrea jecorina. *Proteins Struct. Funct. Bioinforma.* **79**, 2588–2592 (2011).
111. Charavgi, M. D., Dimarogona, M., Topakas, E., Christakopoulos, P. & Chrysina, E. D. The structure of a novel glucuronoyl esterase from Myceliophthora thermophila gives new insights into its role as a potential biocatalyst. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **69**, 63–73 (2013).
112. De Santi, C., Gani, O. A., Helland, R. & Williamson, A. Structural insight into a CE15 esterase from the marine bacterial metagenome. *Sci. Rep.* **7**, 17278 (2017).
113. Agger, J. W., Busk, P. K., Pilgaard, B., Meyer, A. S. & Lange, L. A new functional classification of glucuronoyl esterases by peptide pattern recognition. *Front. Microbiol.* **8**, 1–8 (2017).
114. Rubingh, D. N. Protein engineering from a bioindustrial point of view. *Curr. Opin. Biotechnol.* (1997) doi:10.1016/S0958-1669(97)80062-6.
115. Balabanova, L., Golotin, V., Podvolotskaya, A. & Rasskazov, V. Genetically modified proteins: Functional improvement and chimeragenesis. *Bioengineered* **6**, 262–274 (2015).
116. Ribeiro, L. F. *et al.* Genetically engineered proteins to improve biomass conversion: New advances and challenges for tailoring biocatalysts. *Molecules* **24**, (2019).
117. Axarli, I., Muleta, A. W., Chronopoulou, E. G., Papageorgiou, A. C. & Labrou, N. E. Directed evolution of glutathione transferases towards a selective glutathione-binding site and improved oxidative stability. *Biochim. Biophys. Acta - Gen. Subj.* **1861**, 3416–3428 (2017).
118. Speck, J., Stebel, S. C., Arndt, K. M. & Müller, K. M. Nucleotide Exchange and

Excision Technology DNA Shuffling and Directed Evolution BT - PCR Protocols. in (ed. Park, D. J.) 333–344 (Humana Press, 2011). doi:10.1007/978-1-60761-944-4_24.

119. Lehtonen, S. I. *et al.* Efficient preparation of shuffled DNA libraries through recombination (Gateway) cloning. *Protein Eng. Des. Sel.* **28**, 23–28 (2015).
120. Yang, J., Ruff, A. J., Arlt, M. & Schwaneberg, U. Casting epPCR (cepPCR): A simple random mutagenesis method to generate high quality mutant libraries. *Biotechnol. Bioeng.* **114**, 1921–1927 (2017).
121. Shim, J. *et al.* Improvement of cyclodextrin glucanotransferase as an antistaling enzyme by error-prone PCR. *Protein Eng. Des. Sel.* **17**, 205–211 (2004).
122. Dutta, S., Basak, B., Bhunia, B., Sinha, A. & Dey, A. Approaches towards the enhanced production of Rapamycin by Streptomyces hygroscopicus MTCC 4003 through mutagenesis and optimization of process parameters by Taguchi orthogonal array methodology. *World J. Microbiol. Biotechnol.* **33**, 90 (2017).
123. Windle, C. L. *et al.* Extending enzyme molecular recognition with an expanded amino acid alphabet. *Proc. Natl. Acad. Sci.* **114**, 2610 LP – 2615 (2017).
124. Belsare, K. D. *et al.* Directed evolution of P450cin for mediated electron transfer . *Protein Eng. Des. Sel.* **30**, 119–127 (2017).
125. Zhou, C. *et al.* Second-Generation Engineering of a Thermostable Transketolase (TKGst) for Aliphatic Aldehyde Acceptors with Either Improved or Reversed Stereoselectivity. *ChemBioChem* **18**, 455–459 (2017).
126. Muteeb, G. & Sen, R. Random Mutagenesis Using a Mutator Strain BT - In Vitro Mutagenesis Protocols: Third Edition. in (ed. Braman, J.) 411–419 (Humana Press, 2010). doi:10.1007/978-1-60761-652-8_29.
127. Terao, Y., Miyamoto, K. & Ohta, H. Improvement of the activity of arylmalonate decarboxylase by random mutagenesis. *Appl. Microbiol. Biotechnol.* **73**, 647–653 (2006).
128. Motomitsu, K., Jiesheng, T. & Mamoru, N. Novel Putative Galactose Operon Involving Lacto-N-Biose Phosphorylase in Bifidobacterium longum. *Appl. Environ. Microbiol.* **71**, 3158–3162 (2005).
129. Rigoldi, F., Donini, S., Redaelli, A., Parisini, E. & Gautieri, A. Review: Engineering of thermostable enzymes for industrial applications. *APL Bioeng.* **2**,

(2018).

130. Lehmann, M., Pasamontes, L., Lassen, S. F. & Wyss, M. The consensus concept for thermostability engineering of proteins. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.* **1543**, 408–415 (2000).
131. Reetz, M. T., Carballeira, J. D. & Vogel, A. Iterative saturation mutagenesis on the basis of b factors as a strategy for increasing protein thermostability. *Angew. Chemie - Int. Ed.* **45**, 7745–7751 (2006).
132. Jacak, R., Leaver-Fay, A. & Kuhlman, B. Computational protein design with explicit consideration of surface hydrophobic patches. *Proteins* **80**, 825–838 (2012).
133. Borgo, B. & Havranek, J. J. Automated selection of stabilizing mutations in designed and natural proteins. *Proc. Natl. Acad. Sci.* **109**, 1494 LP – 1499 (2012).
134. Alfarano, P. *et al.* Optimization of designed armadillo repeat proteins by molecular dynamics simulations and NMR spectroscopy. *Protein Sci.* **21**, 1298–1314 (2012).
135. Kilbane, J. J. Biodesulfurization: How to Make it Work? *Arab. J. Sci. Eng.* **42**, 1–9 (2017).
136. Berman, H. M. *et al.* The Protein Data Bank (www.rcsb.org). *Nucleic Acids Res.* (2000) doi:10.1093/nar/28.1.235.
137. Tandon, S., Sharma, A., Singh, S., Sharma, S. & Sarma, S. J. Therapeutic enzymes: Discoveries, production and applications. *J. Drug Deliv. Sci. Technol.* **63**, 102455 (2021).
138. Mohamed Abd El-Aziz, T., Garcia Soares, A. & Stockand, J. D. Snake Venoms in Drug Discovery: Valuable Therapeutic Tools for Life Saving. *Toxins (Basel).* **11**, 564 (2019).
139. Kanda, A., Ono, T., Yoshida, N., Tojo, H. & Okamoto, M. The primary structure of a membrane-associated phospholipase A2 from human spleen. *Biochem. Biophys. Res. Commun.* **163**, 42–48 (1989).
140. Jia, Y. & Villarreal, J. Phospholipases A2 purified from cottonmouth snake venoms display no antibacterial effect against four representative bacterial species. *Toxicon* **151**, 1–4 (2018).

141. Illingworth, S. & Allen, G. Introduction. in *Effective Science Communication (Second Edition)* 1–5 (IOP Publishing, 2020). doi:10.1088/978-0-7503-2520-2ch1.
142. A., E. J. & Jacob, R. Open Access and Global Participation in Science. *Science (80-.).* **323**, 1025 (2009).
143. Viegas, M. *et al.* Biocombustíveis, serão a solução? *Rev. ciênc. elem.* **6**, 25–28 (2018).
144. Pinto, A. V. P. *et al.* Proteínas Virais no Protein Data Bank. *Rev. ciênc. elem.* **8**, (2020).
145. Sousa, S. F. *et al.* Application of quantum mechanics/molecular mechanics methods in the study of enzymatic reaction mechanisms. *Wiley Interdisciplinary Reviews: Computational Molecular Science* (2017) doi:10.1002/wcms.1281.
146. Sousa, S. F., Fernandes, P. A. & Ramos, M. J. Computational enzymatic catalysis – clarifying enzymatic mechanisms with the help of computers. *Phys. Chem. Chem. Phys.* **14**, 12431–12441 (2012).
147. Vlachakis, D., Bencurova, E., Papangelopoulos, N. & Kossida, S. Chapter Seven - Current State-of-the-Art Molecular Dynamics Methods and Applications. in (ed. Donev, R. B. T.-A. in P. C. and S. B.) vol. 94 269–313 (Academic Press, 2014).
148. Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*. (Wiley, 2013).
149. Young, D. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*. (Wiley, 2004).
150. Cerqueira, N. M. F. S. A., Fernandes, P. A. & Ramos, M. J. Protocol for Computational Enzymatic Reactivity Based on Geometry Optimisation. *ChemPhysChem* **19**, 669–689 (2018).
151. Van Der Kamp, M. W. & Mulholland, A. J. Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. *Biochemistry* **52**, 2708–2728 (2013).
152. McPherson, A. Protein Crystallization BT - Protein Crystallography: Methods and Protocols. in (eds. Wlodawer, A., Dauter, Z. & Jaskolski, M.) 17–50 (Springer New York, 2017). doi:10.1007/978-1-4939-7000-1_2.

153. Ziarek, J. J., Baptista, D. & Wagner, G. Recent developments in solution nuclear magnetic resonance (NMR)-based molecular biology. *J. Mol. Med. (Berl).* **96**, 1–8 (2018).
154. Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature* **587**, 157–161 (2020).
155. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **49**, D10–D17 (2021).
156. Muhammed, M. T. & Aki-Yalcin, E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chem. Biol. Drug Des.* **93**, 12–20 (2019).
157. Werner, T., Morris, M. B., Dastmalchi, S. & Church, W. B. Structural modelling and dynamics of proteins for insights into drug interactions. *Adv. Drug Deliv. Rev.* **64**, 323–343 (2012).
158. Cavasotto, C. N. & Phatak, S. S. Homology modeling in drug discovery: current trends and applications. *Drug Discov. Today* **14**, 676–683 (2009).
159. Waterhouse, A. *et al.* SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* (2018) doi:10.1093/nar/gky427.
160. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
161. Levitt, M. Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226**, 507–533 (1992).
162. Katebi, A. R., Kloczkowski, A. & Jernigan, R. L. Structural interpretation of protein-protein interaction network. *BMC Struct. Biol.* **10 Suppl 1**, S4–S4 (2010).
163. Šali, A. & Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
164. Petrey, D. *et al.* Using Multiple Structure Alignments, Fast Model Building, and Energetic Analysis in Fold Recognition and Homology Modeling. *Proteins Struct. Funct. Genet.* **53**, 430–435 (2003).
165. Munsamy, G. & Soliman, M. E. S. Homology Modeling in Drug Discovery - an Update on the Last Decade. *Lett. Drug Des. Discov.* **14**, 1099–1111 (2017).
166. Liu, H. *et al.* Quantum mechanics simulation of protein dynamics on long timescale. *Proteins Struct. Funct. Bioinforma.* **44**, 484–489 (2001).

167. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
168. Chen, V. B. *et al.* MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 12–21 (2010).
169. Wiederstein, M. & Sippl, M. J. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* (2007) doi:10.1093/nar/gkm290.
170. Benkert, P., Künzli, M. & Schwede, T. QMEAN server for protein model quality estimation. *Nucleic Acids Res.* **37**, W510–W514 (2009).
171. Wang, J. *et al.* In Silicon Approach for Discovery of Chemopreventive Agents. *Curr. Pharmacol. Reports* **3**, 184–195 (2017).
172. Moreira, I. S., Fernandes, P. A. & Ramos, M. J. Protein-protein docking dealing with the unknown. *J. Comput. Chem.* **31**, 317–342 (2010).
173. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. J. J. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).
174. de Vries, S. J., van Dijk, M. & Bonvin, A. M. J. J. The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.* **5**, 883–897 (2010).
175. de Vries, S. J. & Bonvin, A. M. J. J. CPORT: A Consensus Interface Predictor and Its Performance in Prediction-Driven Docking with HADDOCK. *PLoS One* **6**, e17695 (2011).
176. Schmitz, C. *et al.* Protein–protein docking with HADDOCK. in *NMR of Biomolecules: Towards Mechanistic Systems Biology* 521–535 (2012). doi:10.1002/9783527644506.
177. van Duin, A. C. T., Dasgupta, S., Lorant, F. & Goddard, W. A. ReaxFF: A Reactive Force Field for Hydrocarbons. *J. Phys. Chem. A* **105**, 9396–9409 (2001).
178. Qiu, Y., Zhong, W., Shao, Y. & Yu, A. Reactive force field molecular dynamics (ReaxFF MD) simulation of coal oxy-fuel combustion. *Powder Technol.* **361**, 337–348 (2020).
179. Case, D. a. *et al.* Amber 12 Reference Manual. *Univ. California, San Fr.* 350

(2012) doi:citeulike-article-id:10779586.

180. D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGra, D. M. Y. and P. A. K. & Case, D. A. Amber 18. *Univ. California, San Fr.* (2018).
181. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
182. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
183. Lemons, D. S. & Gythiel, A. Paul Langevin's 1908 paper "On the Theory of Brownian Motion" ["Sur la théorie du mouvement brownien," C. R. Acad. Sci. (Paris) 146 , 530–533 (1908)]. *Am. J. Phys.* **65**, 1079–1081 (1997).
184. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38**, 3098–3100 (1988).
185. Kohn, W., Becke, A. D. & Parr, R. G. Density Functional Theory of Electronic Structure. *J. Phys. Chem.* **100**, 12974–12980 (1996).
186. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789 (1988).
187. Warshel, A. & Levitt, M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **103**, 227–249 (1976).
188. Fersht, A. R. Profile of Martin Karplus, Michael Levitt, and Arieh Warshel, 2013 Nobel Laureates in Chemistry. *Proc. Natl. Acad. Sci.* **110**, 19656–19657 (2013).
189. Vreven, T. *et al.* Combining Quantum Mechanics Methods with Molecular Mechanics Methods in ONIOM. *J. Chem. Theory Comput.* **2**, 815–826 (2006).
190. Maseras, F. & Morokuma, K. IMOMM: A new integrated ab initio + molecular mechanics geometry optimization scheme of equilibrium structures and transition states. *J. Comput. Chem.* **16**, (1995).
191. Zhou, J. *et al.* QM/MM Studies of the Matrix Metalloproteinase 2 (MMP2)

Inhibition Mechanism of (S)-SB-3CT and its Oxirane Analogue. *J. Chem. Theory Comput.* **6**, 3580–3587 (2010).

192. Xiao, C. & Zhang, Y. Design-atom approach for the quantum mechanical/molecular mechanical covalent boundary: a design-carbon atom with five valence electrons. *J. Chem. Phys.* **127**, 124102 (2007).
193. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gky1049.
194. Vivian, J. P. *et al.* The impact of single cysteine residue mutations on the replication terminator protein. *Biochem. Biophys. Res. Commun.* **310**, 1096–1103 (2003).
195. Longenecker, K. L., Garrard, S. M., Sheffield, P. J. & Derewenda, Z. S. Protein crystallization by rational mutagenesis of surface residues: Lys to Ala mutations promote crystallization of RhoGDI. *Acta Crystallogr. Sect. D* **57**, 679–688 (2001).
196. Olsson, M. H. M., SØndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: Consistent treatment of internal and surface residues in empirical p K a predictions. *J. Chem. Theory Comput.* (2011) doi:10.1021/ct100578z.
197. Martínez-Rosell, G., Giorgino, T. & De Fabritiis, G. PlayMolecule ProteinPrepare: A Web Application for Protein Preparation for Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **57**, 1511–1516 (2017).
198. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).
199. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinforma.* **78**, 1950–1958 (2010).
200. Wang, J. & Hou, T. Application of Molecular Dynamics Simulations in Molecular Property Prediction. 1. Density and Heat of Vaporization. *J. Chem. Theory Comput.* **7**, 2151–2165 (2011).
201. Santos-Martins, D., Calixto, A. R., Fernandes, P. A. & Ramos, M. J. A Buried Water Molecule Influences Reactivity in α-Amylase on a Subnanosecond Time Scale. *ACS Catal.* **8**, 4055–4063 (2018).
202. Calixto, A. R., Ramos, M. J. & Fernandes, P. A. Conformational diversity induces

nanosecond-timescale chemical disorder in the HIV-1 protease reaction pathway. *Chem. Sci.* **10**, 7212–7221 (2019).

203. Ribeiro, A. J. M., Santos-Martins, D., Russo, N., Ramos, M. J. & Fernandes, P. A. Enzymatic Flexibility and Reaction Rate: A QM/MM Study of HIV-1 Protease. *ACS Catal.* **5**, 5617–5626 (2015).
204. Goerigk, L. & Grimme, S. A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions. *Phys. Chem. Chem. Phys.* **13**, 6670–6688 (2011).
205. Listyarini, R. V., Gesto, D. S., Paiva, P., Ramos, M. J. & Fernandes, P. A. Benchmark of density functionals for the calculation of the redox potential of Fe3+/Fe2+ within protein coordination shells. *Front. Chem.* **7**, 1–12 (2019).
206. Sousa, J. P. M., Neves, R. P. P., Sousa, S. F., Ramos, M. J. & Fernandes, P. A. Reaction mechanism and determinants for efficient catalysis by DszB, a key enzyme for crude oil bio-desulfurization. *ACS Catal.* **10**, 9545–9554 (2020).
207. Goerigk, L. *et al.* A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **19**, 32184–32215 (2017).
208. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* (2015) doi:10.1021/acs.jctc.5b00255.
209. Riplinger, C., Pinski, P., Becker, U., Valeev, E. F. & Neese, F. Sparse maps—A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. *J. Chem. Phys.* **144**, 24109 (2016).
210. Paiva, P., Ramos, M. J. & Fernandes, P. A. Assessing the validity of DLPNO-CCSD(T) in the calculation of activation and reaction energies of ubiquitous enzymatic reactions. *J. Comput. Chem.* **41**, 2459–2468 (2020).
211. van Meer, G., Voelker, D. R. & Feigenson, G. W. Membrane lipids: where they are and how they behave. *Nat. Rev. Mol. Cell Biol.* **9**, 112–24 (2008).
212. Fielding, B. A. & Frayn, K. N. Lipid metabolism. *Curr. Opin. Lipidol.* **14**, 389–391 (2003).
213. de Carvalho, C. & Caramujo, M. The Various Roles of Fatty Acids. *Molecules*

23, 2583 (2018).

214. Calder, P. C. Long chain fatty acids and gene expression in inflammation and immunity. *Current Opinion in Clinical Nutrition and Metabolic Care* (2013) doi:10.1097/MCO.0b013e3283620616.
215. Kendall, A. C. *et al.* Distribution of bioactive lipid mediators in human skin. *J. Invest. Dermatol.* (2015) doi:10.1038/jid.2015.41.
216. Bloch, K. & Rittenberg, D. Sources of Acetic Acid in the Animal Body. *J. Biol. Chem* **155**, 243–254.
217. Wakil, S. J., Pugh, E. & Sauer, F. The Mechanism of Fatty Acid Synthesis. *Proc. Natl. Acad. Sci. U. S. A* **52**, 106.
218. Majerus, P. W., Alberts, A. W. & Vagelos, P. R. The Acyl Carrier Protein of Fatty Acid Synthesis: Purification, Physical Properties, and Substrate Binding Site. *Proc. Natl. Acad. Sci.* **51**, 1231–1238 (1964).
219. Majerus, P. W., Alberts, A. W. & Vagelos, P. R. Acyl Carrier Protein, IV. The Identification of 4'-Phosphopantetheine as the Prosthetic Group of the Acyl Carrier Protein. *Proc. Natl. Acad. Sci.* **53**, 410–417 (1965).
220. Smith, S. & Tsai, S.-C. The type I fatty acid and polyketide synthases: a tale of two megasynthases. *Nat. Prod. Rep.* **24**, 1041 (2007).
221. Keatinge-Clay, A. T. The structures of type i polyketide synthases. *Nat. Prod. Rep.* **29**, 1050–1073 (2012).
222. Katsuyama, Y. & Ohnishi, Y. Type III Polyketide Synthases in Microorganisms. in *Methods in Enzymology* vol. 515 359–377 (Elsevier Inc., 2012).
223. Keatinge-Clay, A. T. & Stroud, R. M. The Structure of a Ketoreductase Determines the Organization of the β-Carbon Processing Enzymes of Modular Polyketide Synthases. *Structure* **14**, 737–748 (2006).
224. Jenke-Kodama, H., Sandmann, A., Müller, R. & Dittmann, E. Evolutionary implications of bacterial polyketide synthases. *Mol. Biol. Evol.* **22**, 2027–39 (2005).
225. Neves, R. P. P. *et al.* Engineering of PKS Megaenzymes—A Promising Way to Biosynthesize High-Value Active Molecules. *Top. Catal.* (2021) doi:10.1007/s11244-021-01490-5.
226. Medina, F. E., Neves, R. P. P., Ramos, M. J. & Fernandes, P. A. QM/MM Study

of the Reaction Mechanism of the Dehydratase Domain from Mammalian Fatty Acid Synthase. *ACS Catal.* **8**, 10267–10278 (2018).

227. Medina, F. E., Neves, R. P. P., Ramos, M. J. & Fernandes, P. A. A QM/MM study of the reaction mechanism of human [small beta]-ketoacyl reductase. *Phys. Chem. Chem. Phys.* **19**, 347–355 (2017).
228. Medina, F. E., Ramos, M. J. & Fernandes, P. A. Complexities of the Reaction Mechanisms of Cc Double Bond Reduction in Mammalian Fatty Acid Synthase Studied with Quantum Mechanics/Molecular Mechanics Calculations. *ACS Catal* **9**, 11404–11412.
229. Paiva, P., Sousa, S. F., Ramos, M. J. & Fernandes, P. A. Understanding the Catalytic Machinery and the Reaction Pathway of the Malonyl-Acetyl Transferase Domain of Human Fatty Acid Synthase. *ACS Catal* **8**, 4860–4872.
230. Paiva, P., Sousa, S. F., Fernandes, P. A., Ramos, M. J. & João Ramos, M. Human Fatty Acid Synthase: A Computational Study of the Transfer of the Acyl Moieties from MAT to the ACP Domain. *ChemCatChem* **11**, 3853–3864 (2019).
231. Röttig, M. *et al.* NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39**, W362–W367 (2011).
232. Lin, C. Y. & Smith, S. Properties of the thioesterase component obtained by limited trypsinization of the fatty acid synthetase multienzyme complex. *J. Biol. Chem.* **253**, 1954–62 (1978).
233. Pazirandeh, M., Chirala, S. S., Huang, W. Y. & Wakil, S. J. Characterization of recombinant thioesterase and acyl carrier protein domains of chicken fatty acid synthase expressed in Escherichia coli. *J. Biol. Chem.* **264**, 18195–201 (1989).
234. Libertini, L. J. & Smith, S. Purification and properties of a thioesterase from lactating rat mammary gland which modifies the product specificity of fatty acid synthetase. *J. Biol. Chem.* **253**, 1393–401 (1978).
235. Lenfant, N., Hotelier, T., Bourne, Y., Marchot, P. & Chatonnet, A. Proteins with an alpha/beta hydrolase fold: Relationships between subfamilies in an ever-growing superfamily. *Chem. Biol. Interact.* **203**, 266–8 (2013).
236. Pemble, C. W., Johnson, L. C., Kridel, S. J. & Lowther, W. T. Crystal structure of the thioesterase domain of human fatty acid synthase inhibited by Orlistat. *Nat. Struct. Mol. Biol.* **14**, 704–709 (2007).

237. Zhang, W. *et al.* Crystal structure of FAS thioesterase domain with polyunsaturated fatty acyl adduct and inhibition by dihomo-γ-linolenic acid. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 15757–15762 (2011).
238. Zhang, W. *et al.* Crystal Structure of Fas Thioesterase Domain with Polyunsaturated Fatty Acyl Adduct and Inhibition by Dihomo-Γ-Linolenic Acid. *Proc. Natl. Acad. Sci. U. S. A* **108**, 15757–15762.
239. Buller, A. R. & Townsend, C. A. Intrinsic Evolutionary Constraints on Protease Structure, Enzyme Acylation, and the Identity of the Catalytic Triad. *Proc. Natl. Acad. Sci. U. S. A* **110**, 653– 661.
240. Søndergaard, C. R., Olsson, M. H. M., Rostkowski, M. & Jensen, J. H. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of p K a values. *J. Chem. Theory Comput.* (2011) doi:10.1021/ct200133y.
241. Pemble, C. W., Johnson, L. C., Kridel, S. J. & Lowther, W. T. Crystal Structure of the Thioesterase Domain of Human Fatty Acid Synthase Inhibited by Orlistat. *Nat. Struct. Mol. Biol* **14**, 704–709.
242. Dennington, Roy; Keith, Todd; Millam, J. GaussView 5. *Semichem Inc. , Shawnee Mission, KS* (2009).
243. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general Amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
244. Proietti, G., Abelak, K. K., Bishop-Bailey, D., Macchiarulo, A. & Nobeli, I. Computational modelling of the binding of arachidonic acid to the human monooxygenase CYP2J2. *J. Mol. Model.* **22**, 279 (2016).
245. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins Struct. Funct. Bioinforma.* **65**, 712–725 (2006).
246. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 327–341 (1977) doi:10.1016/0021-9991(77)90098-5.
247. European Commission (EU). The European Green Deal. in vol. COM Annex II

(2019).

248. Katsimpouras, C. *et al.* Enzymatic synthesis of model substrates recognized by glucuronoyl esterases from Podospora anserina and Myceliophthora thermophila. *Appl. Microbiol. Biotechnol.* **98**, 5507–5516 (2014).
249. Rauwerdink, A. & Kazlauskas, R. J. How the Same Core Catalytic Machinery Catalyzes 17 Different Reactions: the Serine-Histidine-Aspartate Catalytic Triad of alpha/beta-Hydrolase Fold Enzymes. *ACS Catal.* **5**, 6153–6176 (2015).
250. d'Errico, C. *et al.* Enzymatic degradation of lignin-carbohydrate complexes (LCCs): Model studies using a fungal glucuronoyl esterase from Cerrena unicolor. *Biotechnol. Bioeng.* **112**, 914–922 (2015).
251. Mosbech, C., Holck, J., Meyer, A. S. & Agger, J. W. The natural catalytic function of CuGE glucuronoyl esterase in hydrolysis of genuine lignin-carbohydrate complexes from birch. *Biotechnol. Biofuels* **11**, 71 (2018).
252. Anna Calarco, A. D. S. A Review on Extremozymes Biocatalysis: A Green Industrial Approach for Biomaterials Production. *J. Biomol. Res. Ther.* **04**, 1–10 (2015).
253. Sarmiento, F., Peralta, R. & Blamey, J. M. Cold and hot extremozymes: Industrial relevance and current trends. *Front. Bioeng. Biotechnol.* **3**, (2015).
254. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* (1997) doi:10.1006/jmbi.1996.0897.
255. Calixto, A. R., Ramos, M. J. & Fernandes, P. A. Influence of Frozen Residues on the Exploration of the PES of Enzyme Reaction Mechanisms. *J. Chem. Theory Comput.* **13**, 5486–5495 (2017).
256. Svensson, M. *et al.* ONIOM: A multilayered integrated MO + MM method for geometry optimizations and single point energy predictions. A test for Diels-Alder reactions and Pt(P(t-Bu)3)2 + H2 oxidative addition. *J. Phys. Chem.* (1996) doi:10.1021/jp962071j.
257. Frisch, M. J. *et al.* Gaussian09 Revision D.01, Gaussian Inc. Wallingford CT. *Gaussian 09 Revision C.01* (2010).
258. Fukui, K. The Path of Chemical Reactions - The IRC Approach. *Acc. Chem. Res.* **14**, 363–368 (1981).

259. Liakos, D. G. & Neese, F. Is It Possible to Obtain Coupled Cluster Quality Energies at near Density Functional Theory Cost? Domain-Based Local Pair Natural Orbital Coupled Cluster vs Modern Density Functional Theory. *J. Chem. Theory Comput.* **11**, 4054–4063 (2015).
260. Chamkin, A. A. & Serkova, E. S. DFT, DLPNO-CCSD(T), and NEVPT2 benchmark study of the reaction between ferrocenium and trimethylphosphine. *J. Comput. Chem.* **41**, 2388–2397 (2020).
261. B:urgi, H. B., Dunitz, J. D., Lehn, J. M. & Wipff, G. Stereochemistry of reaction paths at carbonyl centres. *Tetrahedron* **30**, 1563–1572 (1974).
262. Radisky, E. S. & Koshland, D. E. A clogged gutter mechanism for protease inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 10316–10321 (2002).
263. Weiner, H., Hu, J. H. J. & Sanny, C. G. Rate limiting steps for the esterase and dehydrogenase reaction catalyzed by horse liver aldehyde dehydrogenase. *J. Biol. Chem.* **251**, 3853–3855 (1976).
264. Chin, K. J., Liesack, W. & Janssen, P. H. Opitutus terrae gen. nov., sp. nov., to accommodate novel strains of the division 'Verrucomicrobia' isolated from rice paddy soil. *Int. J. Syst. Evol. Microbiol.* (2001) doi:10.1099/00207713-51-6-1965.
265. Pawlik, A. *et al.* The wood decay fungus Cerrena unicolor adjusts its metabolism to grow on various types of wood and light conditions. *PLoS One* **14**, 1–19 (2019).
266. Morgenstern, I. *et al.* A molecular phylogeny of thermophilic fungi. *Fungal Biol.* (2012) doi:10.1016/j.funbio.2012.01.010.
267. Ernst, H. A. *et al.* The structural basis of fungal glucuronoyl esterase activity on natural substrates. *Nat. Commun.* **11**, 1026 (2020).
268. Watanabe, T. & Koshijima, T. Evidence for an ester linkage between lignin and glucuronic acid in lignin-carbohydrate complexes by DDQ-oxidation. *Agric. Biol. Chem.* (1988) doi:10.1080/00021369.1988.10869116.
269. Hou, L. *et al.* Functional promiscuity correlates with conformational heterogeneity in A-class glutathione S-transferases. *J. Biol. Chem.* **282**, 23264–23274 (2007).
270. Holford, M. Venoms to the rescue. *Science (80-.).* **361**, 842–844.

271. Tasoulis, T. & Isbister, G. K. A Review and Database of Snake Venom Proteomes. *Toxins (Basel).* **9**,.
272. Casewell, N. R. Causes and Consequences of Snake Venom Variation. *Trends Pharmacol. Sci.* **41**, 570–581.
273. Casewell, N. R. Medically important differences in snake venom composition are dictated by distinct postgenomic mechanisms. *Proc Natl Acad Sci U S A* **111**, 9205–10.
274. Massey, D. J. Venom variability and envenoming severity outcomes of the Crotalus scutulatus scutulatus (Mojave rattlesnake) from Southern Arizona. *J Proteomics* **75**, 2576–87.
275. Kalita, B., Mackessy, S. P. & Mukherjee, A. K. Proteomic analysis reveals geographic variation in venom composition of Russell's Viper in the Indian subcontinent: implications for clinical manifestations post-envenomation and antivenom treatment. *Expert Rev Proteomics* **15**, 837–849.
276. Snake-bite envenoming: a priority neglected tropical disease. *Lancet* **390**, 2–2.
277. Keohane, R. O., Lane, M. & Oppenheimer, M. The ethics of scientific communication under uncertainty. *Polit. Philos. Econ.* **13**, 343–368 (2014).
278. Martin, C. & MacDonald, B. H. Using interpersonal communication strategies to encourage science conversations on social media. *PLoS One* **15**, e0241972 (2020).
279. Renata, H., Wang, Z. J. & Arnold, F. H. Expanding the enzyme universe: accessing non-natural reactions by mechanism-guided directed evolution. *Angew. Chemie Int. Ed.* **54**, 3351–3367 (2015).