

DOUTORAMENTO
BIOMEDICINA

Innovative Approach to Fight Tuberculosis and Malaria Targeting the Extraordinary PLP Synthase Macromolecular Complex

André Figueiredo Pina

D

2023





FACULDADE DE MEDICINA
UNIVERSIDADE DO PORTO

ANDRÉ FIGUEIREDO PINA

**INNOVATIVE APPROACH TO FIGHT TUBERCULOSIS AND MALARIA
TARGETING THE EXTRAORDINARY PLP SYNTHASE
MACROMOLECULAR COMPLEX**

Tese de candidatura conducente ao Grau de Doutor em
Biomedicina apresentada à Faculdade de Medicina da
Universidade do Porto

Orientador – Doutor Nuno Manuel Ferreira de Sousa de
Azevedo Cerqueira

Categoria – Investigador

Afiliação – BioSIM Research Group, Faculdade de Medicina
da Universidade do Porto

Co-orientador – Doutor Sérgio Filipe Maia de Sousa

Categoria – Investigador

Afiliação – BioSIM Research Group, Faculdade de Medicina
da Universidade do Porto

INFORMAÇÃO TÉCNICA

TÍTULO

Innovative Approach to Fight Tuberculosis and Malaria Targeting the Extraordinary
PLP Synthase Macromolecular Complex

Tese de candidatura conducente ao Grau de Doutor em Biomedicina submetida à
Faculdade de Medicina da Universidade do Porto

AUTOR

André Figueiredo Pina

DATA

janeiro de 2023

EDITOR: André Figueiredo Pina

CORREIO ELETRÓNICO: andre@andrepina.com

1ª EDIÇÃO: janeiro de 2023

*A todos os que sempre acreditaram em mim....
Em especial a ti Sara, a quem dedico esta tese.*

This PhD thesis was conducted at the UCIBIO@REQUIMTE, in the BioSIM research group of the Faculty of Medicine of the University of Porto under de supervision of Doctor Nuno Manuel Ferreira de Sousa de Azevedo Cerqueira and Doctor Sérgio Filipe Maia de Sousa.

FCT – Fundação para a Ciência e a Tecnologia, I.P., supported this PhD through the studentship SFRH/BD/136594/2018.

The work developed as part of this thesis was financed by national funds from FCT, in the scope of the projects UIDP/04378/2020 and UIDB/04378/2020 of the Applied Molecular Biosciences Unit – UCIBIO and the project LA/P/0140/2020 of the Associate Laboratory Institute for Health and Bioeconomy – i4HB. This work was also partially funded by FCT through the project PTDC/QUI-QFI/31689/2017.

The calculations performed as part of this work were mainly produced using two clusters: the National Distributed Computing Infrastructure – INCD high-performance computing (HPC) cluster funded by FCT and FEDER under the projects 22153–01/SAICT/2016, CPCA/A00/7140/2020, CPCA/A00/7145/2020, and 2021.09782.CPCA; and the FLASH HPC cluster at the Faculty of Medicine of the University of Porto.



This page was intentionally left blank.

AGRADECIMENTOS

Ao concretizar este importante marco no meu percurso académico, não posso deixar de agradecer a todos os que de alguma forma contribuíram para o conseguir alcançar.

Ao meu orientador, Doutor Nuno Cerqueira, manifesto a minha gratidão por tudo o que me ensinou, assim como por todo o apoio e motivação. A ele que me acolheu quando eu não sabia o que fazer do meu futuro, e me abriu as portas para o mundo da Bioquímica Computacional, o meu obrigado! Este agradecimento estende-se ao Doutor Sérgio Sousa, que enquanto meu coorientador sempre me incentivou a ir mais além, acreditando nas minhas capacidades mesmo quando eu não o conseguia fazer. A ele agradeço também a oportunidade de trabalhar nos mais diversos projectos, levando-me a desenvolver um conhecimento mais vasto e profundo sobre esta área. A ambos reforço uma vez mais o meu agradecimento.

A todos os colegas e amigos que ganhei no BioSIM, não sei o que seria de mim sem vós. Muito obrigado por todo o apoio, motivação, e entreajuda que foram a minha “boia de salvamento”, mesmo nos momentos mais difíceis. Ao Henrique, à Juliana, à Rita, à Carla, à Tatiana, ao Fábio e à Cátia, o meu eterno obrigado. Não posso, no entanto, deixar de mencionar o crucial papel do Henrique no meu percurso, a quem para além de um colega e um amigo, considero um mentor. Portanto a ele, um especial obrigado.

Aos meus pais e avós deixo um agradecimento muito especial, por todo o esforço que fizeram pela minha educação. A eles que sempre deram tudo por mim, e me motivaram a ser melhor, muito obrigado.

A ti Joana, minha querida irmã e ex-colega de casa, obrigado por nunca me deixares desistir. És o meu porto seguro, com quem sei sempre poder contar. A ti Rodrigo, obrigado por seres o apoio da minha irmã mesmo quando eu não o consigo ser. Obrigado, ainda por todas as nossas conversas e discussões, mais ninguém teria tanta paciência como tu para aturar o meu constante debitar de informação.

Aos meus sogros, Mário e Manuela, a quem deixo também um agradecimento especial por todo o apoio.

Aos meus eternos amigos, Fábio e Chico, que apesar da distância e da minha ausência, fazem sempre parecer como que se nenhum tempo tivesse passado. Obrigado por todas as aventuras passadas e por todas aquelas que ainda estão para vir.

Quero também deixar um especial obrigado à minha psicóloga, a Dr.^a Marta Magalhães Basto, por toda a ajuda, em especial nesta fase final.

E por fim, se há alguém a quem devo ter conseguido levar esta tese até ao fim és tu Sara. Não existem palavras suficientes para descrever tudo o que fizeste por mim, não só nestes últimos meses, mas ao longo de todos estes anos ao teu lado. Obrigado por toda a paciência (MUITA PACIÊNCIA), atenção e encorajamento sem os quais eu sei que não teria conseguido. A ti, Sara, dedico esta tese.

ABSTRACT

The work presented in this thesis is divided into two main stages.

In the first stage, we explored the catalytic mechanisms of Pdx2 (EC 4.3.3.6) and XometC (EC 4.4.1.1) enzymes. Pdx2 is the glutaminase domain of pyridoxal 5'-phosphate (PLP) synthase, a macromolecular complex regarded as an important drug target in the fight against tuberculosis and malaria. XometC, on the other hand, is an essential PLP-dependent enzyme from a pathogenic bacterium that affects rice production worldwide, thus being regarded as a potential target for new bactericides. In this thesis, their catalytic mechanisms were studied using hybrid quantum mechanics/molecular mechanics (QM/MM) computational methods. These methods provide an atomistic level of information on the unknown mechanisms of these enzymes that, together with Gibbs free energy profiles, provide important clues on how the catalytic process takes place, highlighting key steps, particularly the rate-limiting step that controls the overall kinetics of the catalytic process. The obtained results agree very well with the available experimental data, and the generated knowledge can be used to develop new drugs targeting these enzymes, providing new solutions for the pharmaceutical and agricultural industries.

The second stage of this thesis is focused on DNA, particularly non-canonical DNA molecules, mainly non-B DNA conformations and DNA aptamers. Under physiological conditions, DNA is primarily found in its common B-DNA form. However, it was found that specific nucleotide sequences are more prone to form non-B DNA conformations, which have been linked with mutagenic instability, mainly in mitochondrial DNA (mtDNA). One example is Structure A, a stretch of the control region of mtDNA, predicted to form a non-B DNA conformation, implicated in the impairment of the mitochondrial replicative process. Therefore, its three-dimensional (3D) structure was predicted, along with other mtDNA stretches of the same length, and their stability was evaluated using molecular dynamics (MD) simulations. The results have shown that when compared to the 3D structure predicted for the other mtDNA stretches evaluated, Structure A is more stable. Regarding DNA aptamers, the goal was to identify, from a set of potential new aptamers, which had a higher affinity for cathepsin B, a protease that is overexpressed in several cancers. A combination of 3D structure prediction, MD simulations, protein-DNA molecular docking, and binding free energy calculations was used to accomplish this task. The results highlight the increasing role of computational methods in this highly emerging field of research of DNA aptamers.

Keywords: pyridoxal 5'-phosphate (PLP), PLP biosynthesis, Pdx2, PLP-dependent enzymes, XometC, catalytic mechanism, QM/MM, non-B DNA conformations, DNA aptamers, molecular dynamics, molecular docking, binding free energy calculations

RESUMO

O trabalho apresentado nesta tese está dividido em duas secções.

Na primeira, explorámos os mecanismos catalíticos das enzimas Pdx2 (EC 4.3.3.6) e XometC (EC 4.4.1.1). A Pdx2 é o domínio glutaminase da piridoxal 5'-fosfato (PLP) sintase, um complexo macromolecular considerado um alvo terapêutico importante na luta contra a tuberculose e a malária. A XometC é uma enzima dependente de PLP essencial para uma bactéria patogénica que afeta a produção de arroz a nível mundial, sendo considerada um potencial alvo para o desenvolvimento de novos bactericidas. Os seus mecanismos catalíticos foram estudados utilizando métodos computacionais híbridos de mecânica quântica/mecânica molecular (QM/MM) que nos permitem obter informação com detalhe atómico. Estes dados, aliados aos perfis de energia livre de Gibbs obtidos, permitem perceber como o processo catalítico ocorre, assim como desvendar estados intermediários importantes que controlam a cinética das enzimas em questão. Estes dados estão de acordo com os resultados experimentais disponíveis, permitindo obter uma versão consensual dos seus mecanismos catalíticos que pode ser usada para o desenvolvimento de novos fármacos e bactericidas para as indústrias farmacêutica e agroalimentar, respetivamente.

A segunda secção foca-se no ADN, em particular em conformações de ADN não-B e aptâmeros de ADN. Apesar de em condições fisiológicas, o ADN encontrar-se principalmente na forma B, verificou-se que certas sequências nucleotídicas são mais propensas a formar conformações de ADN não-B. Estas têm sido relacionadas com instabilidade mutagénica, principalmente no ADN mitocondrial (ADNmt). Um exemplo é a Estrutura A, uma porção da região de controlo do ADNmt implicada no comprometimento do processo de replicação mitocondrial. Portanto, a sua estrutura tridimensional (3D) foi prevista, juntamente com a de outras porções de ADNmt, sendo a sua estabilidade avaliada utilizando simulações de dinâmica molecular (MD). Os resultados demonstraram que, quando comparada com a estrutura 3D prevista para os outros ADNmt avaliados, a estrutura A é mais estável. Relativamente aos aptâmeros de ADN, o objetivo era identificar, partindo de um conjunto de potenciais novos aptâmeros, aqueles que tinham uma maior afinidade para a cathepsina B, uma protease sobre-expressa em vários cancros. Para realizar esta tarefa foram utilizados métodos como, previsão da estrutura 3D, simulações MD, acoplamento molecular de proteína-ADN, e cálculos de energia livre de ligação. Os resultados destacam o papel crescente dos métodos computacionais na investigação de aptâmeros de ADN.

Palavras-chave: piridoxal 5'-fosfato (PLP), biossíntese de PLP, Pdx2, enzimas dependentes de PLP, XometC, mecanismo catalítico, QM/MM, conformações de ADN não-B, aptâmeros de ADN, dinâmica molecular, acoplamento molecular, cálculos de energia livre de ligação

PREAMBLE

This thesis is organized into six main chapters:

A. INTRODUCTION

Chapter A provides a general introduction to the two main sections of this PhD project: proteins and DNA. The proteins section focuses on the pyridoxal 5'-phosphate (PLP) cofactor, covering topics such as its biosynthesis, the pathways involved, and the enzymes that require it to catalyse reactions. The chemistry of PLP is also discussed, exploring how a single cofactor can facilitate the catalysis of so many different chemical reactions. The DNA section covers topics like the existence of various conformations of DNA, focusing primarily on non-B DNA conformations, and their implications in disease. Moreover, the use of DNA molecules, known as aptamers, as a targeted therapeutic approach is also presented.

B. METHODOLOGY

Chapter B provides a review of the theoretical principles of the main computational methodologies employed in the development of this PhD project. This chapter is divided into five main sections:

1. Preparation of the model system
2. Molecular mechanics
3. Quantum mechanics
4. Hybrid methods
5. Enzymatic kinetics

Furthermore, this chapter also explains some of the choices made regarding the methods used in the execution of this work.

C. RESULTS AND DISCUSSION

Chapter C, the most relevant chapter of this thesis, is divided into four main sections, the first two devoted to proteins and the others to DNA:

1. PLP biosynthesis
2. PLP-dependent enzymes
3. Non-B DNA conformations
4. DNA aptamers

Section 1 focuses on the study of the catalytic mechanism of Pdx2—the glutaminase subunit of PLP synthase—a key enzyme in PLP biosynthesis. The key aspects that differ between its non-canonical Cys-His-Glu catalytic triad and the well-known Ser-His-Asp/Glu catalytic triad are also discussed.

Section 2 describes the catalytic mechanism behind the L-serine ammonia-lyase activity of XometC, a promiscuous PLP-dependent enzyme found in a bacteria responsible for the devastation of rice crops worldwide.

Section 3 focuses on the study of Structure A, a non-B DNA conformation that naturally occurs in the mitochondrial DNA, implicated as the cause of mutagenic instability. Therefore, a workflow was implemented to predict the 3D structure and study the dynamics of non-B DNA conformations.

Finally, section 4 is dedicated to the study of the interaction of several DNA aptamers with Cathepsin B, a protein over-expressed in some cancers. Their affinity toward their target was ranked, allowing the *in silico* prediction of the best ones, which was ultimately experimentally validated.

Each section is composed of an initial summary of the work, followed by a complete portrayal of the resulting published scientific article.

D. FINAL REMARKS

Chapter D summarizes the key aspects of the work developed throughout this PhD project and provides a general conclusion of its main contributions to the scientific community.

E. REFERENCES

Chapter E gathers all the bibliographic references mentioned throughout this thesis.

F. APPENDIX

Chapter F contains all the appendixes that complement what was presented throughout the main content of this thesis, being cross-referenced whenever appropriate.

This page was intentionally left blank.

All scientific publications, oral communications, and posters, result of this PhD project, are presented below in the following list.

List of Scientific Publications (7)

Reviews (1)

1. Rocha, J. F., **Pina, A. F.**, Sousa, S. F., & Cerqueira, N. M. F. S. A. (2019). PLP-dependent enzymes as important biocatalysts for the pharmaceutical, chemical and food industries: A structural and mechanistic perspective. *Catalysis Science & Technology*, 9(18), 4864–4876. doi: [10.1039/C9CY01210A](https://doi.org/10.1039/C9CY01210A) [HOT ARTICLE]

Articles (4)

1. **Pina, A. F.**, Sousa, S. F., & Cerqueira, N. M. F. S. A. (2021). The Catalytic Mechanism of Pdx2 Glutaminase Driven by a Cys-His-Glu Triad: A Computational Study. *ChemBioChem*, 23(9), e202100555. doi: [10.1002/cbic.202100555](https://doi.org/10.1002/cbic.202100555)
2. Pereira, A. C.[‡], **Pina, A. F.**[‡], Sousa, D.[‡], Ferreira, D., Santos-Pereira, C., Rodrigues, J. L., Melo, L. D. R., Sales, G., Sousa, S. F., & Rodrigues, L. R. (2022). Identification of novel aptamers targeting cathepsin B-overexpressing prostate cancer cells. *Molecular Systems Design & Engineering*, 7(6), 637–650. doi: [10.1039/D2ME00022A](https://doi.org/10.1039/D2ME00022A)
3. **Pina, A. F.**, Sousa, S. F., Azevedo, L., & Carneiro, J. (2022). Non-B DNA conformations analysis through molecular dynamics simulations. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1866(12), 130252. doi: [10.1016/j.bbagen.2022.130252](https://doi.org/10.1016/j.bbagen.2022.130252)
4. **Pina, A. F.**, Fernandes, H. S., Sousa, S. F., & Cerqueira, N. M. F. S. A. (2022). Catalytic mechanism of XometC, a promiscuous PLP-dependent enzyme with L-serine ammonia-lyase activity. [Under submission]

[‡] The authors contributed equally to the work.

Proceedings (2)

1. **Pina, A. F.**, Sousa, S. F., & Cerqueira, N. M. F. S. A. (2021). A computational study on the catalytic mechanism of Pdx2: a glutaminase containing the Cys-His-Glu triad. *Paper presented at the Proceedings of MOL2NET'21, Conference on Molecular, Biomedical & Computational Sciences and Engineering, 7th edition*
2. **Pina, A. F.**, Sousa, S. F., & Cerqueira, N. M. F. S. A. (2022). A computational study on the catalytic mechanism of pyridoxal 5'-phosphate synthase. *FEBS Open Bio*, 12(S1), 227. doi: [10.1002/2211-5463.13440](https://doi.org/10.1002/2211-5463.13440)

List of Communications (7)

Posters (6)

1. **Pina, A. F.**, Sousa, S. F., & Cerqueira, N. M. F. S. A. (2019). Application of molecular dynamics simulations in the study of the PLP synthase macrocomplex. *BImBS school 2019 – Bioinformatics meets BioSimulations in protein and DNA studies: from theory to practice*, Lugano, Switzerland (October 5–12) (**Appendix**: page 263)
2. **Pina, A. F.**, Sousa, S. F., & Carneiro, J. (2021). 3D prediction of non-B DNA conformations associated with mtDNA genomic instability. *10th Bioinformatics Open Days*, online due to COVID-19 restrictions (May 5–7) (**Appendix**: page 264)
3. **Pina, A. F.**, Sousa, S. F., & Cerqueira, N. M. F. S. A. (2021). A computational study on the catalytic mechanism of Pdx2: a glutaminase containing the Cys-His-Glu triad. *VIII EJIBCE – Encontro De Jovens Investigadores Em Biologia Computacional Estrutural*, Coimbra, Portugal (December 20) (**Appendix**: page 265)
4. **Pina, A. F.**, Sousa, S. F., & Cerqueira, N. M. F. S. A. (2022). A computational study on the catalytic mechanism of pyridoxal 5'-phosphate synthase. *2022 IUBMB–FEBS–PABMB Young Scientists' Forum*, Vimeiro, Portugal (July 6–9) (**Appendix**: page 266)
5. **Pina, A. F.**, Sousa, S. F., & Cerqueira, N. M. F. S. A. (2022). A computational study on the catalytic mechanism of pyridoxal 5'-phosphate synthase. *The Biochemistry Global Summit*, Lisboa, Portugal (July 9–14) (**Appendix**: page 266)
6. **Pina, A. F.**, & Sousa, S. F. (2022). Development of computational methods for aptamer design for biomedical applications. *5th Meeting in Medicinal Biotechnology /*

2nd Iberian Congress in Medicinal Biotechnology, Porto, Portugal (October 21)
(**Appendix:** page 267)

Oral Communications (1)

1. **Pina, A. F.**, Sousa, S. F., & Cerqueira, N. M. F. S. A. (2022). The catalytic mechanism of Pdx2 glutaminase: a QM/MM approach. *11th Bioinformatics Open Days*, Braga, Portugal (March 3–5)

This page was intentionally left blank.

INDEX

AGRADECIMENTOS	iii
ABSTRACT	v
RESUMO	vii
PREAMBLE	ix
<i>List of Scientific Publications (7)</i>	<i>xiii</i>
Reviews (1)	xiii
Articles (4)	xiii
Proceedings (2).....	xiv
<i>List of Communications (7)</i>	<i>xiv</i>
Posters (6)	xiv
Oral Communications (1)	xv
INDEX	xvii
INDEX OF FIGURES	xxv
INDEX OF EQUATIONS	xxxv
INDEX OF TABLES	xli
LIST OF ABBREVIATIONS	xliii
A. INTRODUCTION	1
1. <i>Central Dogma of Molecular Biology</i>	3
2. <i>Proteins</i>	5
2.1. Pyridoxal 5'-phosphate (PLP): an essential cofactor.....	5
2.2. Biosynthesis of PLP	7
2.2.1. De novo pathways	8
2.2.1.1. DXP-dependent pathway.....	8
2.2.1.2. DXP-independent pathway.....	8
a. Pdx2	10
b. Pdx1	13
2.2.2. Salvage pathway	14
2.3. PLP-dependent enzymes	15
2.3.1. Versatility of the enzymes that employ PLP as a cofactor	17

2.3.2. Common aspects of the catalytic process of PLP-dependent enzymes...	19
3. DNA.....	23
3.1. Non-B DNA conformations: a cause of mutagenic instability	25
3.2. DNA aptamers: a promising therapeutic tool	27
B. METHODOLOGY	29
1. Preparation of the model system	32
2. Molecular Mechanics	35
2.1. Force Fields	35
2.1.1. Bonded Interactions	36
2.1.2. Non-Bonded Interactions	38
2.2. Energy Minimization	39
2.3. Molecular Dynamics Simulations	40
2.3.1. Integration Time Step.....	41
2.3.2. Ensembles	42
2.3.3. Temperature and Pressure Control.....	43
2.3.4. Periodic Boundary Conditions	43
2.3.5. Non-Bonded Interactions Cutoff.....	44
2.3.6. Analysis of the Molecular Dynamics Simulation	44
2.4. Molecular Docking.....	45
2.4.1. Search algorithms	46
2.4.2. Scoring Functions	46
2.4.3. HADDOCK.....	47
2.5. Binding Free Energy Calculations	48
3. Quantum Mechanics.....	51
3.1. Energy Quantization	52
3.2. Wave-Particle Duality	53
3.3. Heisenberg Uncertainty Principle	54
3.4. The Schrödinger Equation.....	54
3.4.1. Born's Statistical Interpretation of the Wavefunction	56
3.4.2. The Hamiltonian Operator.....	56
3.5. Variational Principle.....	57
3.6. Born-Oppenheimer Approximation	58
3.7. Basis Sets	59
3.7.1. Slater and Gaussian Type Orbitals	60
3.7.2. Contracted Basis Sets	61
3.7.2.1. Pople style basis sets	61

a. Diffuse and polarization functions	62
3.7.2.2. Ahlrichs basis sets.....	62
3.7.2.3. Dunning's correlation consistent basis sets	63
a. Basis set extrapolation.....	63
3.8. Electronic Hamiltonian.....	64
3.8.1. Hartree-Fock.....	64
3.8.1.1. Many-Electrons System.....	65
3.8.1.2. Hartree-Fock Theory	65
3.8.1.3. Pauli Exclusion Principle and Slater Determinants	66
3.8.1.4. Hartree-Fock Limit: Electron Correlation.....	66
3.8.2. Post-Hartree-Fock Methods	67
3.8.2.1. Configuration Interaction (CI).....	67
3.8.2.2. Multiconfiguration Self-Consistent Field Methods	69
3.8.2.3. Møller-Plesset Perturbation Theory	69
3.8.2.4. Coupled Cluster (CC)	71
a. Domain-Based Local Pair Natural Orbital (DLPNO) Coupled-Cluster Method	72
3.8.3. Density Functional Theory	72
3.8.3.1. Hohenberg-Kohn Theorems	73
3.8.3.2. Kohn-Sham Equations.....	74
3.8.3.3. Exchange-Correlation Density Functionals	75
a. Local Density Approximation (LDA)	78
b. Generalized Gradient Approximation (GGA).....	78
c. Hybrid Functionals	79
d. Composite Methods.....	79
3.8.3.4. Dispersion effects.....	80
3.8.4. Semi-empirical Methods.....	81
3.8.5. Potential Energy Surface (PES).....	82
3.8.5.1. Locating Stationary Points on the PES	83
a. Locating Transition Structures	84
3.8.5.2. Characterizing Stationary Points	84
3.8.5.3. Potential Energy Surface Scans	86
3.9. Nuclear Hamiltonian	86
3.9.1. Zero-point Energy Calculation.....	87
3.9.2. Translational, rotational, and vibrational approximations	87
3.9.3. Gibbs Free Energy.....	88
4. Hybrid Methods.....	91

4.1. Quantum Mechanics/Molecular Mechanics	91
4.1.1. Additive and Subtractive Methods.....	92
4.1.2. Mechanical and Electrostatic Embedding.....	93
4.1.2.1. Mechanical Embedding	93
4.1.2.2. Electrostatic Embedding	94
4.1.3. Boundary Conditions.....	94
5. <i>Enzymatic Kinetics</i>	97
C. RESULTS AND DISCUSSION	101
1. <i>PLP biosynthesis</i>	105
1.1. Catalytic mechanism of the Pdx2 glutaminase subunit of PLP synthase	107
1.1.1. Introduction	109
1.1.2. Methodology	111
1.1.2.1. Model preparation	111
1.1.2.2. Structure minimization	112
1.1.2.3. QM/MM	112
1.1.3. Results and Discussion.....	115
1.1.3.1. Stage 1: Activation of Cys ₈₇	115
1.1.3.2. Stage 2: Formation of the glutamyl-thioester intermediate.....	117
1.1.3.3. Stage 3: Hydrolysis of the glutamyl-thioester intermediate	121
1.1.3.4. Stage 4: Enzymatic turnover.....	122
1.1.3.5. Energy Profile.....	123
1.1.4. Conclusion.....	125
2. <i>PLP-dependent enzymes</i>	129
2.1. Catalytic mechanism of XometC – a β -elimination reaction	131
2.1.1. Introduction	132
2.1.2. Methodology	135
2.1.2.1. Model preparation	135
2.1.2.2. Structure minimization	135
2.1.2.3. QM/MM	136
2.1.2.4. Simulating UV-Vis spectra.....	139
2.1.3. Results and Discussion.....	139
2.1.3.1. Stage 1: Formation of external aldimine with L-serine	140
2.1.3.2. Stage 2: β -Elimination	142
2.1.3.3. Stage 3: Formation of the internal aldimine	145
2.1.3.4. Stage 4: Reprotonation of Tyr ₁₁₂ and enzymatic turnover	149
2.1.3.5. Non-enzymatic catalysis: Conversion of 2-aminoacrylate into pyruvate	149

2.1.3.6. UV-VIS	152
2.1.4. Conclusion	153
3. <i>Non-B DNA conformations</i>	159
3.1. Analysis of the 3D conformation of Structure A – a non-B DNA conformation associated with mtDNA mutagenic instability	161
3.1.1. Introduction	163
3.1.2. Methodology	164
3.1.2.1. Determination of secondary structures	164
3.1.2.2. UNAFold predictions accuracy	165
3.1.2.3. Generation of three-dimensional non-B DNA conformations	166
3.1.2.4. Molecular dynamics simulations of the 3D non-B DNA conformations	167
3.1.2.5. Analysis of the molecular dynamics simulations	168
3.1.2.6. Root-Mean-Square Deviation (RMSD), end-to-end Distance (R_{ee}), and radius of gyration (R_g)	169
3.1.2.7. Hairpin states and H-bond trajectory analysis	169
3.1.2.8. Cluster analysis	169
3.1.2.9. Conservation scores analysis	170
3.1.3. Results and Discussion	170
3.1.3.1. UNAFold predictions accuracy	170
3.1.3.2. Structure A is stable throughout molecular dynamics simulations	171
3.1.3.3. Structure A cluster analysis	174
3.1.3.4. Human, primate and mammalian conservation scores and RMSF variation	176
3.1.3.5. Stack regions present higher conservation scores and low RMSF variation	178
3.1.3.6. Somatic and germline disease-associated variants within the Structure A elements	180
3.1.4. Conclusion	181
4. <i>DNA aptamers</i>	183
4.1. Identification of new aptamers targeting cathepsin B – a protease overexpressed in prostate cancer	185
4.1.1. Introduction	187
4.1.2. Methodology	189
4.1.2.1. Reagents	189
4.1.2.2. Preparation of CatB-conjugated magnetic beads	189
a. Magnetic beads surface activation	189
b. CatB immobilization	190
4.1.2.3. SELEX procedure	190
4.1.2.4. Amplification and ssDNA pool purification	191

4.1.2.5. Next-generation sequencing and data analysis	191
4.1.2.6. Generation of three-dimensional DNA aptamers models	191
4.1.2.7. Molecular dynamics simulations of the aptamers.....	193
4.1.2.8. Aptamer-CatB Docking	193
4.1.2.9. Molecular dynamics simulations of the docked CatB–aptamer complexes	194
4.1.2.10. MM-GBSA method for binding free energies calculations	194
4.1.2.11. Aptamers' characterization	195
4.1.2.12. Cell-Binding assays	196
4.1.2.13. Statistical analysis	196
4.1.3. Results and Discussion.....	197
4.1.3.1. Selection of aptamers against CatB aided by NGS sequencing.....	197
4.1.3.2. Analysis of aptamers MD simulation	199
4.1.3.3. CatB-Aptamers molecular docking	201
4.1.3.4. Refinement of CatB-Aptamer complexes and binding free energy calculations	202
4.1.3.5. Calculation of the aptamer's affinity properties	207
4.1.3.6. Binding assays of the selected aptamers against CatB-overexpressing prostate cancer cells	209
4.1.4. Conclusion	210
D. FINAL REMARKS	211
1. <i>Proteins</i>	213
1.1. Pdx2 – a key enzyme in PLP biosynthesis	213
1.2. XometC – role of PLP as a cofactor.....	215
2. <i>DNA</i>	219
3. <i>Methodology</i>	221
3.1. Model preparation	221
3.2. Molecular mechanics.....	221
3.3. Quantum mechanics	223
E. REFERENCES	227
F. APPENDIX.....	259
1. <i>Posters</i>	261
1.1. Application of molecular dynamics simulations in the study of the PLP synthase macrocomplex.....	263
1.2. 3D prediction of non-B DNA conformations associated with mtDNA genomic instability	264

1.3. A computational study on the catalytic mechanism of Pdx2: a glutaminase containing the Cys-His-Glu triad.....	265
1.4. A computational study on the catalytic mechanism of pyridoxal 5'-phosphate synthase.....	266
1.5. Development of computational methods for aptamer design for biomedical applications	267
2. <i>Covers</i>	269
2.1. Cover of the ChemBioChem.....	271
3. <i>Reviews</i>	273

This page was intentionally left blank.

INDEX OF FIGURES

Figure 1. Structure of all B ₆ vitamers, pyridoxine (PN), pyridoxal (PL), pyridoxamine (PM), pyridoxine 5'-phosphate (PNP), pyridoxal 5'-phosphate (PLP; the active form of vitamin B ₆), and pyridoxamine 5'-phosphate (PMP).....	6
Figure 2. DXP-dependent and DXP-independent PLP biosynthetic pathways. Epd, erythrose 4-phosphate dehydrogenase; PdxB, 4-phosphoerythronate dehydrogenase; SerC, 3-phosphoserine aminotransferase; PdxA, 4-phosphohydroxy-L-threonine dehydrogenase; PdxJ, PNP synthase; DXPS, 1-deoxyxylulose 5-phosphate synthase; PNPOx, PNP oxidase; Pdx2 (glutaminase subunit) and Pdx1 (synthase subunit) of the PLP synthase complex.....	7
Figure 3. 3D representation of the PLP synthase macro-complex. (A) Top view. (B) Side view. The Pdx1 synthase subunits (represented in blue) assemble in a dodecameric form, two hexamers on top of each other, with one Pdx2 glutaminase subunit (represented in orange) bounded to each of them.....	9
Figure 4. General catalytic mechanism employed by enzymes with Ser-His-Asp/Glu or Cys-His-Asp/Glu catalytic triads. At first glance, independently of the nucleophile used by the enzyme, the mechanism is very similar. However, there are some nuances, as we shall see below, especially regarding the activation of the nucleophile. Adapted from (50).	12
Figure 5. Biosynthetic salvage pathway for the “recycling” of B ₆ vitamers into PLP. PLK, pyridoxal kinase; PNPOx, pyridoxine/pyridoxamine 5'-phosphate oxidase.	14
Figure 6. Distribution of all known PLP-dependent enzymes with an EC number attributed across the different enzyme groups.....	15
Figure 7. Absorption spectra of the PLP-dependent Serine palmitoyltransferase (SPT), a key enzyme in sphingolipid biosynthesis, adapted from the work of Ikushiro et al. (62). SPT catalyzes the decarboxylative condensation of L-serine with palmitoyl-CoA to generate 3-ketodihydrosphingosine. First, L-serine reacts with the internal aldimine (IA), forming the external aldimine (EA), then only if palmitoyl-CoA is present does the reaction continue with the formation of the quinonoid intermediate (QI), and ultimately the product. This was discovered by UV-visible spectroscopy using an analogue to palmitoyl-CoA that is not able to react with L-serine but whose presence triggers the formation of the QI, as observed in this absorption spectra. Furthermore, increasing concentrations of the	

analogue lead to a higher absorption at the typical QI wavelength, indicative of a higher QI formation. 16

Figure 8. Overview of the different types of reactions catalyzed by PLP-dependent enzymes. According to the Dunathan hypothesis, the bond to be cleaved is the one perpendicular to the plane of the pyridinium ring of PLP. Depending on the enzyme, this leads to the formation of one of three types of quinonoid intermediates. These intermediates can then undergo several different reactions, demonstrating the high versatility of PLP-dependent enzymes. 18

Figure 9. Schematic representation of the general mechanism for the activation of a PLP-dependent enzyme with the formation of the internal (IA) and its transamination with a substrate, forming the external (EA) aldimines. 22

Figure 10. Nucleotides comprise a nucleobase, a pentose sugar (deoxyribose in the case of DNA), and a phosphate group. Carbon residues in the pentose are numbered 1' to 5'. New nucleotides are added at the 3' hydroxyl group by covalently bonding to the 5' phosphate group of the incoming nucleotide. Bases can be divided into two categories: purines and pyrimidines. Adapted from (80). 23

Figure 11. Schematic representation of several of the most common non-B DNA conformations. Adapted from (88). 25

Figure 12. Illustration of all the bonded interactions considered in MM force fields used in the study of biomolecular systems. 38

Figure 13. Categorization of the exchange-correlation density functionals according to Perdew's "Jacob's ladder". Adapted from (228). 77

Figure 14. Model of a 3D potential energy surface showcasing, among other things, minima (a reactant and product), a transition state structure, and a reaction path. Adapted from (256). 82

Figure 15. Saddle point at the surface of the saddle-shaped function: $f_{x,y} = 12kx^2 - y^2$. Adapted from (115). 83

Figure 16. Analysis and visualization of the vibrational frequencies associated with a TS performed with molUP, a VMD plugin. If the structure has only one imaginary frequency (verified on the panel shown on the left), and the atom displacements observed

correspond to the reaction coordinate of interest (represented by the red vectors), we can validate the structure as a real TS. Adapted from the original molUP article (257). 85

Figure 17. Representation of a QM/MM model. All the surrounded area (protein + water molecules) is treated with a MM method, while the residues within the highlighted area are treated with a QM method. 92

Figure 18. Schematic representation of a computationally determined energy profile. As observed the $\Delta G^\ddagger$ of the rate-limiting step (shown in orange) is identical to the energy barrier determined by the Eyring equation for the experimental k_{cat} (red). The Gibbs free energy of reaction, ΔGR , of the rate-limiting step is also shown in blue. 97

Figure 19. General reaction catalyzed by Pdx2. The ammonia produced by Pdx2 is the source of the ring nitrogen of PLP, denoted by a red arrow in the image. 110

Figure 20. (A) Top and (B) Side view of the PLP synthase macrocomplex represented in QuickSurf. The Pdx1 and Pdx2 subunits are coloured in cyan and orange, respectively. One of the Pdx1:Pdx2 dimers is represented in NewCartoon, with the substrates, ribose 5-phosphate and glutamine, represented in VDW. 111

Figure 21. (Left) Structure of the dimer Pdx1:Pdx2 (cyan and orange, respectively) and Licorice representation of the HL layer. (Right) The substrate, glutamine, and active site residues are represented in Licorice and coloured in orange and blue, respectively. 113

Figure 22. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 1. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. Glutamine, the substrate, is represented in orange, whereas the rest is in light blue. The yellow vectors illustrate the major vibrational modes corresponding to the imaginary frequency of the TS (shown in the bottom right corner). Relevant interatomic distances (in Å) for **R** **TS** **P** are shown, as well as the activation and reaction Gibbs free energies for this step. 116

Figure 23. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 2. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. Glutamine, the substrate, is represented in orange, whereas the rest is in light blue. The yellow vectors illustrate the major vibrational modes corresponding to the imaginary frequency of the TS (shown in the bottom right corner). Relevant interatomic

distances (in Å) for **R** **TS** **P** are shown, as well as the activation and reaction Gibbs free energies for this step. 118

Figure 24. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 3. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. Glutamine, the substrate, is represented in orange, whereas the rest is in light blue. The yellow vectors illustrate the major vibrational modes corresponding to the imaginary frequency of the TS (shown in the bottom right corner). Relevant interatomic distances (in Å) for **R** **TS** **P** are shown, as well as the activation and reaction Gibbs free energies for this step. 119

Figure 25. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 4. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. Glutamine, the substrate, is represented in orange, whereas the rest is in light blue. The yellow vectors illustrate the major vibrational modes corresponding to the imaginary frequency of the TS (shown in the bottom right corner). Relevant interatomic distances (in Å) for **R** **TS** **P** are shown, as well as the activation and reaction Gibbs free energies for this step. 120

Figure 26. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 5. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. Glutamine, the substrate, is represented in orange, whereas the rest is in light blue. The yellow vectors illustrate the major vibrational modes corresponding to the imaginary frequency of the TS (shown in the bottom right corner). Relevant interatomic distances (in Å) for **R** **TS** **P** are shown, as well as the activation and reaction Gibbs free energies for this step. 121

Figure 27. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 6. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. Glutamine, the substrate, is represented in orange, whereas the rest is in light blue. The yellow vectors illustrate the major vibrational modes corresponding to the imaginary frequency of the TS (shown in the bottom right corner). Relevant interatomic distances (in Å) for **R** **TS** **P** are shown, as well as the activation and reaction Gibbs free energies for this step. 123

Figure 28. Energy profile determined for the catalysis of the conversion of glutamine into glutamate by Pdx2, with the release of ammonia. The values in parentheses correspond to the corresponding activation or reaction Gibbs free energy of each step. All energy values were computed with DLPNO-CCSD(T)/CBS//B3LYP/6-31G(d,p):ff14SB..... 124

Figure 29. Overview of the catalytic mechanism of Pdx2, with the indication of the distinct phases: **Activation & Enzymatic Turnover**, **Deamination**, and **Hydrolysis**. The Gibbs free energies of activation ($\Delta G^\ddagger$) and reaction (ΔG_R) are expressed in kcal.mol⁻¹..... 126

Figure 30. Schematic representation of the catalytic mechanism of XometC, a promiscuous PLP-dependent enzyme able to catalyze the β -elimination of the hydroxyl group of L-serine with the formation of 2-aminoacrylate, later converted into pyruvate in the cytosol. The formation of the internal and external aldimines is also depicted. 134

Figure 31. (Left) Structure of the XometC homotetramer, with each subunit coloured differently. The residues belonging to the QM region of the bottom right subunit are represented in Licorice. (Right) PLP, the substrate, and the zwitterionic L-serine are coloured pink, while the active site residues are coloured blue. The residues Tyr₅₈, Arg₆₀, and Thr₂₀₉ belong to the adjacent subunit. Each molecule is represented in Licorice and is identified by a circular label. 137

Figure 32. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 1. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. PLP and L-Ser are represented in pink, while the rest are in light blue. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown. 140

Figure 33. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of steps 2-4. The structures of the reactants and products are represented in a wedge-dash manner. The structures of the TS are represented three-dimensionally, using the Licorice drawing method of VMD. PLP and L-Ser are represented in pink, while the rest are in light blue. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown. 142

Figure 34. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of steps 5-6. The structures of the reactants and products are represented in a wedge-dash manner. The structures of the TS are represented three-dimensionally, using the Licorice drawing method of VMD. PLP and L-Ser/2AA are represented in pink, while the rest are in light

blue. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown. 144

Figure 35. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 7. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. PLP and 2AA are represented in pink, while the rest are in light blue. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown. 145

Figure 36. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 8. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. PLP and 2AA are represented in pink, while the rest are in light blue. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown. 146

Figure 37. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of steps 9-13. The structures of the reactants and products are represented in a wedge-dash manner. The structures of the TS are represented three-dimensionally, using the Licorice drawing method of VMD. PLP and 2AA are represented in pink, while the rest are in light blue. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown. 149

Figure 38. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 14. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. PLP and 2AA are represented in pink, while the rest are in light blue. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown. 149

Figure 39. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of the steps involved in the non-enzymatic catalysis of the deamination of 2AA into pyruvate that occurs in the cytosol. The structures of the reactants and products are represented in a wedge-dash manner. The structures of the TS are represented three-dimensionally, using the Licorice drawing method of VMD, and are coloured in pink. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown. 151

Figure 40. UV-Vis absorption spectra calculated with TD-DFT of four important intermediates of the catalytic mechanism of XometC: the internal aldimine (IA), the external aldimine with L-serine (EAS), the quinonoid (QUI), and the external aldimine with 2-aminoacrylate (EAA). The electrostatic potential was calculated and mapped on the surface of the reaction intermediates..... 153

Figure 41. Overview of the catalytic mechanism of the PLP-dependent enzyme XometC uncovered in this work..... 155

Figure 42. Energy profile determined for the catalysis of L-serine conversion into 2-aminoacrylate by XometC. All energy values were computed with DLPNO-CCSD(T)/CBS//r²SCAN-3c/mTZVPP:ff14SB. 156

Figure 43. Energy profile determined for the non-enzymatic catalysis of the deamination of 2-aminoacrylate to pyruvate. All energy values were computed with DLPNO-CCSD(T)/CBS. 157

Figure 44. Analysis of the molecular dynamics simulation of the 3D Structure A conformation and controls. Graphs A, B, and C are represented as a function of time for the 100 ns of simulation. (A) RMSD values. All heavy atoms were used in the calculation of the mass-weighted RMSD. (B) Radius of gyration. (C) Distance between the centre of mass of the 3' and 5' ends of the DNA conformations. (D) Box-and-whiskers plot of the number of hydrogen bonds established during the molecular dynamics simulation. The whiskers represent the minimum and maximum values, while the box extends from the 25th to 75th percentiles, with the line in the middle plotted at the median value. 171

Figure 45. R_g and R_{ee} space of each structure is represented as a heatmap. Each frame of the MD simulations corresponds to one point in the graph, located by its (R_g , R_{ee}) value. The heat (i.e., colour) corresponds to the local density of points, with red being the most densely populated and blue the less densely populated. The R_g - R_{ee} space of Structure A is zoomed in to highlight the presence of two distinct populations. Representative structures of each of these populations are shown on the right. The main difference between the two populations is the nature of the interaction between T1 and A93 (i.e., 5' and 3' residues, respectively). The terminal residues of the structures belonging to the population with a higher R_{ee} establish Watson-Crick base pairing, whereas the ones with a lower R_{ee} perform van der Waals stacking interactions. 173

Figure 46. Superimposition of the representative structure of the major cluster obtained from the molecular dynamics simulation with the initial 3D prediction. The clustering was

performed using the K-means method, and 5 clusters were obtained. The structures from the molecular dynamics simulations are coloured according to residue RMSD by comparison with the predicted structure. 175

Figure 47. Analysis of the conservation scores of the 3D Structure A conformation considering the human intraspecific variation (Human Alignment) and the interspecific variation in Primates (Primates Alignment) and Mammals (Mammals Alignment). The horizontal axis shows the mtDNA nucleotide position, the primary vertical axis shows the conservation score percentage (%), and the secondary vertical axis shows the RMSFs values (Å). The red line represents the RMSF variation along the mtDNA structure A region. 177

Figure 48. Analysis of the PPI conservation scores of the 3D Structure A conformation considering the structural element inside Structure A and the human intraspecific variation (Human Alignment), and the interspecific variation in Primates (Primates Alignment) and Mammals (Mammals Alignment). The horizontal axis shows the structural element, and the primary vertical axis shows the conservation score percentage (%). 179

Figure 49. Analysis of the RMSFs values (Å) of the 3D Structure A conformation considering the structural element inside Structure A. The horizontal axis shows the structural element, and the vertical axis represents the RMSF values (Å). 180

Figure 50. Summary of the workflow used to generate the DNA aptamer models, from their nucleotide sequence to their 3D structure. fw, forward; rc, reverse complement; UNAFold, unified nucleic acid folding, and hybridization package. 192

Figure 51. Primary structure and phylogenetic tree analysis. The random region (50 nt) of the ten top-ranked oligonucleotide sequences, along with their respective reverse complementary sequences obtained after NGS, were evaluated by performing (A) multiple sequence alignments on Geneious 9.1.4 using a MAFFT alignment with the FFT-NS-2 algorithm (% of similarity), and (B) phylogenetic analysis generated using the Geneious tree builder employing the Tamura–Nei genetic distance model and the neighbour-joining method. 198

Figure 52. Flowchart of the workflow applied in this work starting from the generated 3D models of the DNA aptamers. fw, forward; rc, reverse complement; MD, molecular dynamics; HADDOCK, high ambiguity driven protein–protein DOCKing; CatB, cathepsin B; MM-GBSA, molecular mechanics-generalized born surface area. 200

Figure 53. Comparison of the Haddock score and total binding free energy (in kcal.mol⁻¹) obtained using the MM-GBSA method for the 20 aptamers under study. The aptamer that was tested experimentally is highlighted with a grey circle (Apt08rc). 204

Figure 54. Analysis of the molecular dynamics simulation of the CatB-Apt08rc complex. (A) Representative structure of the CatB-Apt08rc complex obtained by K-means clustering of the molecular dynamics simulation observed from the side (left) and from below (right). CatB is represented in surface and coloured according to its residue type. Blue, positively charged side groups; red, negatively charged side groups; white, non-polar with aliphatic or aromatic side groups; Grey, polar with uncharged side groups. (B) Representative structure of the CatB-Apt08rc complex with CatB represented in NewCartoon, with all the residues within 3.5 Å from the Apt08rc represented in Licorice and coloured according to its residue type. (C) Pairwise RMSD of the trajectory of the CatB-Apt08rc complex. The colour scale is represented on the right going from a dark purple (0 Å) to yellow (30 Å). (D) RMSD representation of either the entire complex, Apt08rc, or CatB as a function of time for the 100 ns of the simulation..... 206

Figure 55. Hydrogen bonds and salt bridges established between CatB and Apt08rc. (A) Number of hydrogen bonds established between the CatB and Apt08rc, as a function of time for the 100 ns of the MD simulation. (B) Number of salt bridges established between CatB and Apt08rc, as a function of time for the 100 ns of MD simulation..... 207

Figure 56. Calibration curve (A) and kinetic assay (B) for Apt08rc, and calibration curve (C) and kinetic assay (D) for Apt01. Experiments were conducted in duplicate, and results correspond to the mean ± standard deviation..... 208

Figure 57. Binding assays with PC-3 cells (cellular density of 1×10⁴ cells per well) and increased concentrations of library (grey), Apt01 (red), and Apt08rc (blue) at room temperature. ****p < 0.0001..... 209

Figure 58. Schematic representation of each protocol with the respective duration time for each reaction step. This excludes any exploration of the reactional space by performing multiple scans. 225

This page was intentionally left blank.

INDEX OF EQUATIONS

Equation 1. General force field equation representing the total energy of a system treated with MM.....	36
Equation 2. Potential energy function for the bonded interactions term of a general biomolecular force field, consisting of bond stretch, bond angle bend, and dihedral torsion. Some force fields also include a fourth term for improper dihedral torsion.....	36
Equation 3. Energy function for the bond term. kl : force constant, l : bond length, l_0 : equilibrium bond length.	37
Equation 4. Energy function for the angle term. $k\theta$: force constant, θ : bond angle, θ_0 : equilibrium bond angle.	37
Equation 5. Energy function for the dihedral term, which is described by an N-term truncated Fourier series. V_n : dihedral barrier height, n : periodicity, ϕ : dihedral angle, γ_n : dihedral phase.....	37
Equation 6. Energy function for the improper dihedral term. $k\psi$: force constant, ψ : improper torsional angle, ψ_0 : equilibrium improper torsional angle.....	38
Equation 7. Potential energy function for the non-bonded interactions term of the general force field. i, j : atom pair, ϵ_{ij} : potential well depth, σ_{ij} : interatomic distance at the energy minimum, r_{ij} : interatomic distance of atoms i and j , q_i and q_j : point charge of the atoms i and j , respectively, ϵ_0 : electric constant.	38
Equation 8. Mathematical relation between the force $\mathbf{F}_i$ acting upon particle i , and its potential energy U . $U(\mathbf{r}_1, \dots, \mathbf{r}_N)$ represents the potential energy of N interacting atoms as a function of their positions $\mathbf{r}_i = x_i, y_i, z_i$	40
Equation 9. Second order derivative of Newton's second law. $\mathbf{F}_i$: force acting on particle i , m_i : mass of the particle i , $\mathbf{r}_i$: position vector of particle i , t : time.....	40
Equation 10. Basic equation of the Verlet algorithm derived from the Taylor expansions for the positions $\mathbf{r}_i(t)$. This equation is accurate to an order given by the fourth power of the time step Δt	41

- Equation 11.** HADDOCK scoring function. E_{vdW} : van der Waals intermolecular energy, E_{elec} : electrostatic intermolecular energy, E_{desolv} : desolvation energy, E_{AIR} : distance restraints energy.48
- Equation 12.** The binding free energy (ΔG_{bind}) is estimated from the free energies of the complex (PL), protein (P), and ligand (L). However, when using the 1-trajectory protocol, only the complex is simulated. The ensemble average of the free receptor and ligand is created simply by removing the respective atoms.49
- Equation 13.** The free energy is calculated by combining the so-called gas phase energy contributions (i.e., E_{bnd} : bonded interactions energy, E_{el} : electrostatic interactions energy, and E_{vdW} : van der Waals interactions energy) with the solvation free energy components (G_{pol} : polar contribution to the solvation free energy, and G_{np} : non-polar contribution to the solvation free energy). In addition, entropy contributions may be added as a further refinement (TS).49
- Equation 14.** Relation between the energy of a photon (E_{photon}) and its frequency (ν). h : Planck's constant.53
- Equation 15.** De Broglie's equation. λ : wavelength; h : Planck's constant; p : momentum of the particle; m : mass of the particle; v : particle's velocity.53
- Equation 16.** Heisenberg's Uncertainty Principle. Δx : uncertainty in position; Δp : uncertainty in momentum; $\hbar$: reduced Planck constant, which is equal to the Planck constant h divided by 2π54
- Equation 17.** Three-dimensional time-dependent Schrödinger equation.....55
- Equation 18.** Laplacian operator.....55
- Equation 19.** One-dimensional, time-dependent Schrödinger equation.55
- Equation 20.** One-dimensional, time-independent Schrödinger equation.55
- Equation 21.** Reduced Schrödinger equation.55
- Equation 22.** The Hamiltonian operator H is the sum of the kinetic energy operator T , and the potential energy operator V56
- Equation 23.** Decomposition of the Hamiltonian operator (H) into its different kinetic and potential energy terms. T_e : electrons' kinetic energy; T_N : nuclei's kinetic energy; V_{Ne} :

nuclei-electron attraction; V_{ee} : electron-electron repulsion; V_{NN} : nuclei-nuclei repulsion.
56

Equation 24. Differentiation of the energy terms of the Hamiltonian in atomic units for simplicity. KE : kinetic energy; N : number of electrons; M : number of nuclei; M_A : mass of nucleus A ; Z : atomic number; r_{iA} : distance between electron i and nucleus A ; r_{ij} : distance between electrons i and j ; R_{AB} : distance between nuclei A and B . ∇^2 is the Laplace operator and is equal to $\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$57

Equation 25. The Variational Principle states that the energy of any approximate ground state wave function is always greater than or equal to the exact energy of the ground state. E_0 : exact ground state energy; E_ϕ : approximate ground state energy.....57

Equation 26. The energy of the trial wavefunction can be calculated as the expectation value of the Hamiltonian (numerator) divided by the normalization integral (denominator). E_ϕ : approximate ground state energy; ψ_ϕ^* : complex conjugate of the approximate ground state wavefunction; ψ_ϕ : approximate ground state wavefunction; H : Hamiltonian acting on ψ_ϕ58

Equation 27. Division of the Hamiltonian operator (H) into electronic (H_{elec}) and nuclear (H_{nuc}) Hamiltonians.....58

Equation 28. Electronic Hamiltonian (H_{elec}).58

Equation 29. Nuclear Hamiltonian (H_{nucl}).59

Equation 30. Representation of an orbital as the sum of K basis functions. χ_i : orbital i ; K : number of basis functions; $C_{\mu i}$: coefficient; χ_μ : basis function.59

Equation 31. STO equation. N : normalization constant; $Y_{l,m}$: spherical harmonic functions that depend on the angular momentum quantum number l and the magnetic quantum number m ; n : principle quantum number; ζ : constant related to the effective charge of the nucleus.60

Equation 32. GTO equation. N : normalization constant; $Y_{l,m}$: spherical harmonic functions that depend on the angular momentum quantum number l and the magnetic quantum number m ; n : principle quantum number; ζ : constant related to the effective charge of the nucleus.60

Equation 33. Wavefunction as a product of individual electron wavefunctions.....65

Equation 34. One-electron spin-orbital.	65
Equation 35. Wavefunction as a product of N spin-orbital functions.	65
Equation 36. Slater determinant for an N -electron system. χ_x : spin-orbital where x denotes the position and spin of a single electron.	66
Equation 37. Exact ground-state electronic wavefunction Ψ where the c s represent expansion coefficients and the limits in the summation indices ($a < b$, etc.) ensure that a given excited determinant only appears once in the summation.	68
Equation 38. Zero-order Hamiltonian (H_{HF}) given by the sum of N_e one-electron Fock operators (f_i), where N_e is the number of electrons in the system.	70
Equation 39. Perturbation contribution. H_{elec} : electronic Hamiltonian.	70
Equation 40. Perturbation contribution for electron i . i and j : electrons; m : occupied molecular orbitals; J_m : Coulomb operator; K_m : exchange operator.	70
Equation 41. According to the CC method, the exact electronic wavefunction Ψ is a product of the exponential operator eC , and the HF wavefunction Ψ_0	71
Equation 42. Exponential operator eC defined by a series expansion.	71
Equation 43. Cluster operator C	71
Equation 44. Hypothetical functional Ffx	73
Equation 45. Electron density at the nucleus. ρ : spherically averaged density; Z : atomic number.	73
Equation 46. Ground-state energy E_0 as the functional of the electron density ρ_0	74
Equation 47. Second theorem of Hohenberg-Kohn.	74
Equation 48. Hamiltonian for the reference system.	74
Equation 49. Kohn-Sham orbitals.	75
Equation 50. Exchange-correlation energy.	75
Equation 51. Exchange-correlation energy of the B3LYP functional. A , B , and C are empirical parameters.	79

Equation 52. Zero-point energy (ZPE) correction.	87
Equation 53. Energy at temperature T	87
Equation 54. Frequency of a harmonic oscillator. μ : reduced mass; k : force constant....	88
Equation 55. Enthalpy contribution estimated by thermal corrections. k_B : Boltzmann's constant; T : temperature.	88
Equation 56. Entropy contribution estimated by thermal corrections.....	88
Equation 57. Gibbs free energy.....	89
Equation 58. Energy function for the QM/MM additive method. EQM : energy of the part of the system calculated with QM; EMM : energy of the part of the system calculated with MM; $EQM/MMInteraction$: interaction energy between the QM and MM parts.....	92
Equation 59. Energy function for the QM/MM subtractive method. $EQMModel$: energy of the model part calculated with QM; $EMMAll$: energy of the entire system calculated with MM; $EMMModel$: energy of the model part calculated with MM.....	93
Equation 60. Arrhenius equation. k : reaction rate constant; A : Arrhenius constant; E_a : activation energy of the reaction; R : ideal gas constant; T : temperature.....	98
Equation 61. Eyring equation. k : reaction rate constant; κ : transmission coefficient; k_B : Boltzmann constant; T : temperature; h : Planck's constant; $\Delta G^\ddagger$: Gibbs free energy of activation; R : ideal gas constant.	98
Equation 62. $\Delta G_{bind} = \Delta G_{complex} - \Delta G_{CatB} + \Delta G_{Apt}$	195
Equation 63. $\Delta G = \Delta E_{gas} + \Delta G_{sol} - T\Delta S$	195
Equation 64. $\Delta E_{gas} = \Delta E_{int} + \Delta E_{ELE} + \Delta E_{VDW}$	195
Equation 65. $\Delta G_{sol} = \Delta G_{GB} + \Delta G_{Surf}$	195
Equation 66. $y = start + [(end - start) \times x^n / k^n + x^n]$	196

This page was intentionally left blank.

INDEX OF TABLES

Table 1. Accuracy results of UNAFold to predict hairpin structures using a database of experimentally determined NMR structures of 22 single-stranded DNA with hairpins.	166
Table 2. Summary of the results of the molecular dynamics simulations, excluding the first 10 ns. The results are represented as Mean (SD).	172
Table 3. Summary of the Pearson's correlations, considering the conservation scores and the molecular dynamics RMSF of Structure A.	177
Table 4. Descriptive statistics of Structure A structural elements analysis, considering the conservation scores and the molecular dynamics RMSF.	179
Table 5. Summary of the Pearson's correlations, considering the conservation scores, the disease-associated mutations, and somatic mutations.	181
Table 6. Top 10 sequences and their complementary sequences (bold) with the highest number of copies retrieved from the total number of sequences obtained by next-generation sequencing (NGS).	197
Table 7. Haddock results and binding free energy calculations. The energies were calculated using 500 structures of the last 50 ns of the MD simulations of the CatB-Apt docked complexes using the MM-GBSA method. The polar contribution to the solvation free energy (ΔG_{GB}) and the electrostatic energy (ΔE_{ELE}) are accounted in the polar ΔG , while the non-polar ΔG refers to the non-polar contribution to the solvation free energy (ΔG_{Surf}) plus the van der Waals energy (ΔE_{VDW}). Haddock score is provided in arbitrary units (a. u.), buried surface area in $\text{\AA}^2$, and the energy values in $\text{kcal.mol}^{-1} \pm$ standard deviation. The aptamer chosen for experimental validation is highlighted in bold.	201
Table 8. Intra- and intermolecular interactions of each aptamer. The number of base pairs, Watson–Crick (W.C.) hydrogen bonds, intermolecular hydrogen bonds, intermolecular salt bridges, and buried surface area were calculated for the last 50 ns of the MD simulations of the CatB-Apt docked complexes. The base pairs, hydrogen bonds, and salt bridges were presented as mean $\pm$ SD. The buried surface area is presented in $\text{\AA}^2 \pm$ standard deviation (SD). The aptamer chosen for experimental validation is highlighted in bold.	203

This page was intentionally left blank.

LIST OF ABBREVIATIONS

2AA	2-Aminoacrylate
3D	Three-Dimensional
ADN	Ácido Desoxirribonucleico
ADNmt	Ácido Desoxirribonucleico Mitocondrial
AIR	Ambiguous Interaction Restraint
AMBER	Assisted Model Building with Energy Refinement
ANOVA	Analysis of Variance
ASP	Astex Statistical Potential
ATP	Adenosine Triphosphate
B3LYP	Becke, Three-Parameter, Lee-Yang-Parr
BB	Bacterial Blight
BD	Bürgi-Dunitz
BSSE	Basis Set Superposition Error
CASSCF	Complete Active Space Self-Consistent Field
catB	Cathepsin B
CBS	Complete Basis Set
CC	Coupled Cluster
CCD	Coupled Cluster Doubles
CCSD	Coupled Cluster Singles and Doubles
CCSD(T)	Coupled Cluster Single, Double, And Perturbative Inclusion of Triple Excitations
CEA	Carcinoembryonic Antigen
cGTO	Contracted Gaussian Type Orbital
CI	Configuration Interaction
CLL	Chronic Lymphocytic Leukemia
CPU	Central Processing Unit
CSF	Configuration State Function
DCI	Doubly Excited CI
DDEC	Density Derived Electrostatic and Chemical
DFT	Density Functional Theory
DFTB	Density Functional Tight Binding
DLPNO	Domain-Based Local Pair Natural Orbital
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic Acid
DXP	Deoxyxylulose 5-Phosphate
DXPS	Deoxyxylulose 5-Phosphate Synthase
EA	External Aldimine
EAA	External Aldimine With 2-Aminoacrylate
EAS	External Aldimine With L-Serine
EC	Enzyme Commission
EDC	1-Ethyl-3-(3-Dimethylaminopropyl)Carbodiimide
EDTA	Ethylenediaminetetraacetic Acid
EM	Electron Microscopy
EpCAM	Epithelial Cell Adhesion Molecule
FBS	Fetal Bovine Serum

FCC	Fraction of Common Contacts
FL	Flippin-Lodge
FMN	Flavin Mononucleotide
fw	Forward
GAFF	General AMBER Force Field
GATase	Glutamine Amidotransferase
GB	Generalized Born
GBSA	Generalized Born Surface Area
GBW	Geometry-Basis-Wavefunction
gCP	Geometrical Counter-Poise Correction
GGA	Generalized Gradient Approximation
GOLD	Genetic Optimisation for Ligand Docking
GPU	Graphics Processing Unit
GROMACS	Groningen Machine for Chemical Simulations
GS1	Gem-Diamine Reaction Intermediate 1
GTO	Gaussian Type Orbitals
HADDOCK	High Ambiguity Driven Protein-Protein Docking
HEG	Homogeneous Electron Gas
HF	Hartree-Fock
HL	High-Level Layer
HMR	Hydrogen Mass Repartitioning
HPC	High-Performance Computing
HPLC	High-Performance Liquid Chromatography
HS	HADDOCK Score
HVR-I	Hypervariable Region I
IA	Internal Aldimine
IRC	Intrinsic Reaction Coordinate
JASP	Jeffreys's Amazing Statistics Program
KE	Kinetic Energy
LAMMPS	Large-Scale Atomic/Molecular Massively Parallel Simulator
L-BFGS	Limited-Memory Broyden-Fletcher-Goldfarb-Shanno Algorithm
LCAO	Linear Combination of Atomic Orbitals
LCPO	Linear Combinations of Pairwise Overlaps
LDA	Local Density Approximation
LINCS	Linear Constraint Solver
LL	Low-Level Layer
LR	Low Responders
LSDA	Local Spin Density Approximation
LYP	Lee-Yang-Parr Correlation Functional
MAFFT	Multiple Alignment using Fast Fourier Transform
MBPT	Many-Body Perturbation Theory
<i>Mbt</i>	<i>Mycobacterium tuberculosis</i>
MCSCF	Multiconfiguration Self-Consistent Field
MD	Molecular Dynamics
MEP	Minimal Energy Path
MES	2-(N-Morpholino)Ethanesulfonic Acid
MFE	Minimum Free Energy
mGGA	Meta-Generalized Gradient Approximation

MIDD	Maternally Inherited Diabetes and Deafness
MM	Molecular Mechanics
MOE	Molecular Operating Environment
MP	Møller-Plesset
MPPT	Møller-Plesset Perturbation Theory
M ^{pro}	SARS-CoV-2 Main Protease
mRNA	Messenger Ribonucleic Acid
mtDNA	Mitochondrial Deoxyribonucleic Acid
MUC1	Mucin-1
NAC	Net Atomic Charge
NADH	Reduced Nicotinamide Adenine Dinucleotide
NADPH	Nicotinamide Adenine Dinucleotide Phosphate
NAMD	Nanoscale Molecular Dynamics
NCBI	National Center for Biotechnology Information
NEB	Nudged Elastic Band
NGS	Next-Generation Sequencing
NHS	N-Hydroxysuccinimide
NMR	Nuclear Magnetic Resonance
NPC	Nuclear Pore Complex
nt	Nucleotide
OMP	Orotidine Monophosphate
ONIOM	Our Own N-Layered Integrated Molecular Orbital and Molecular Mechanics
OPLS	Optimized Potentials for Liquid Simulations
PB	Poisson-Boltzmann
PBC	Periodic Boundary Conditions
PBE	Perdew-Burke-Enzerhof
PBS	Phosphate Buffered Saline
PBSA	Poisson-Boltzmann Surface Area
PC-3	Prostate Cancer Cell Line
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PES	Potential Energy Surface
<i>Pf</i>	<i>Plasmodium falciparum</i>
PIS	Percentage of Identical Sites
PL	Pyridoxal
PLK	Pyridoxal Kinase
PLP	Pyridoxal 5'-Phosphate
PLP	Piecewise Linear Potential scoring function
PLPS	Pyridoxal 5'-Phosphate Synthase
PM	Pyridoxamine
PMEMD	Particle Mesh Ewald Molecular Dynamics
PMP	Pyridoxamine 5'-Phosphate
PN	Pyridoxine
PNP	Pyridoxine 5'-Phosphate
PNPOx	Pyridoxine/Pyridoxamine 5'-Phosphate Oxidase
PPI	Percentage of Pairwise Identity
PSMA	Prostate-Specific Membrane Antigen

PT	Perturbation Theory
QI	Quinonoid Intermediate
QM	Quantum Mechanics
QUI	Quinonoid
QZ	Quadruple-Zeta
R5P	Ribose 5-Phosphate
rc	Reverse Complement
RDF	Radial Distribution Function
R _{ee}	End-to-End Distance
RESP	Restrained Electrostatic Potential
RFO	Rational Function Optimization
R _g	Radius of Gyration
RMSD	Root-Mean-Square Deviation
RMSF	Root-Mean-Square Fluctuation
RNA	Ribonucleic Acid
SASA	Solvent-Accessible Surface Area
SCF	Self-Consistent Field
SD	Standard Deviation
SDCI	Singly and Doubly Excited CI
SELEX	Systematic Evolution of Ligands by Exponential Enrichment
SLTCAP	Screening Layer Tally by Container Average Potential
SMD	Solvation Model Based on Density
SP	Single Point
SPC/E	Extended Simple Point Charge
SPT	Serine Palmitoyltransferase
SQM	Semiempirical Quantum Mechanics
ssDNA	Single-Stranded Deoxyribonucleic Acid
ssRNA	Single-Stranded Ribonucleic Acid
STO	Slater Type Orbitals
TDA	Tamm-Dancoff Approximation
TD-DFT	Time-Dependent Density Functional Theory
TIP3P	Transferable Intermolecular Potential 3P
TPSS	Tao-Perdew-Staroverov-Scuseria
tRNA	Transfer Ribonucleic Acid
TS	Transition State
TZ	Triple-Zeta
UMP	Uridine Monophosphate
UNAFold	Unified Nucleic Acid Folding and Hybridisation Package
UV	Ultraviolet
VDW	Van der Waals
VMD	Visual Molecular Dynamics
WB	Washing Buffer
XC	Exchange-Correlation
Xoo	<i>Xanthomonas Oryzae Pv. oryzae</i>
ZPE	Zero-Point Energy

A. INTRODUCTION

This page was intentionally left blank.

This first chapter provides an overview of the two main topics that will be the focus of this PhD thesis: proteins (particularly enzymes) and DNA — the underlying molecules of the central dogma of molecular biology.

Since the primary goal of this PhD project was to study the chemistry of pyridoxal 5'-phosphate (PLP) biosynthesis carried out by the PLP synthase macromolecular complex, a particular focus will be placed on this cofactor, as well as its biosynthetic pathways and related enzymes. Furthermore, the role played by PLP as a cofactor in several different enzymes will also be addressed.

Regarding DNA, the focus will be split between non-B DNA conformations and their contribution to mutagenic instability, and DNA aptamers, which are molecules with remarkable potential as recognition elements and drug delivery systems.

Moreover, the importance of computational simulations as an essential tool for understanding the structure and function of proteins and DNA will also be highlighted.

1. Central Dogma of Molecular Biology

Understanding how life works and operates intricately has always been in the spotlight of scientific knowledge. The Central Dogma of Molecular Biology is a fundamental concept in molecular biology that describes the flow of genetic information. Francis Crick first proposed it in 1958 (1) and further reiterated in 1970 in a Nature paper (2) that “once information has passed into protein, it cannot get out again”. According to the central dogma, information can flow from nucleic acid to nucleic acid or from nucleic acid to protein, but not from protein to protein or from protein to nucleic acid.

This concept has since become a cornerstone of our understanding of inheritance and gene expression mechanisms. It also helps to explain the role of nucleic acids, like DNA and RNA, and proteins in the cell and the mechanisms by which mutations in DNA can lead to changes in an organism's phenotype.

DNA (deoxyribonucleic acid) is a long polymer that encodes the genetic information of an organism. In eukaryotes, it can be found primarily in the nucleus, condensed in the form of linear chromosomes, but also in the mitochondria, in the form of circular chromosomes. Although it usually exists in the form of a double helix, especially as B-DNA *in vivo*, it can also exist in a single-stranded form that can acquire non-B DNA conformations. This last point will be the focus of **section 3.1**. Moreover, the unique molecular recognition properties of DNA, together with its high stability, make DNA an obvious choice in the

development of aptamers, i.e., short synthetic oligonucleotides created to bind to a specific target molecule. For this reason, they are often considered antibody mimetics with several biomedical applications. A more in-depth description of DNA aptamers can be found in **section 3.2**.

As a result of the transcription and translation processes, cells are able to synthesize proteins. Protein biosynthesis is a core biological process involved in the dynamic maintenance of a cellular proteome to ensure protein homeostasis (also known as proteostasis). Proteins are responsible for several critical functions inside and outside cells, either as enzymes, structural proteins, hormones, and antibodies, among many others. They are undoubtedly the workhorses of any living organism.

In summary, the Central Dogma of Molecular Biology is a fundamental concept that provides a framework for understanding the flow of genetic information and the role of nucleic acids and proteins in the cell. In the following section, the subject of proteins is explored in more detail, giving special attention to the enzymes involved in the biosynthesis of the cofactor pyridoxal 5'-phosphate (PLP) (**section 2.2**), as well as those that require it, as a cofactor, for their catalytic activity (**section 2.3**).

2. Proteins

Proteins are large biomolecules that result from messenger RNA (mRNA) translation into long chains of amino acid residues by the ribosome. Their structure ranges from a single polypeptide chain (i.e., with a single chain) to some of the largest multimeric complexes currently known, such as the Nuclear Pore Complex (NPC), built from ~1000 copies of ~34 different nucleoporins (3,4). As mentioned, protein function ranges from providing structure to cell signalling and ligand binding (e.g., antibodies and hormones). Nevertheless, some of the cell's best-known and most essential proteins are enzymes.

Enzymes are biological catalysts, i.e., they increase the rate of chemical reactions that otherwise would be very slow or never occur in a human timescale. For example, the decarboxylation of orotidine monophosphate (OMP) into uridine monophosphate (UMP), an essential reaction in the *de novo* biosynthesis of pyrimidine nucleotides, would take 78 million years if uncatalyzed. However, when catalyzed by the enzyme OMP decarboxylase, the reaction takes only 18 milliseconds, accelerating the reaction rate by a factor of 10^{17} (5).

Nevertheless, despite most enzymes relying only on the correct positioning of the necessary amino acid residues within the active site for the reaction to occur, which results from a specific protein folding, others need a little more “help” — a cofactor.

Cofactors are chemical compounds required by enzymes to help them fulfil their role as catalysts. They can be divided into two categories: organic cofactors, also known as coenzymes, or inorganic cofactors, such as the metal ions Mg^{2+} , Cu^+ and Zn^{2+} . While some cofactors are produced in our bodies, like NADH and NADPH, others must be acquired through our diet. One good example is vitamin B₆, the precursor of pyridoxal 5'-phosphate (PLP).

2.1. Pyridoxal 5'-phosphate (PLP): an essential cofactor

Vitamins are organic molecules required by an organism to maintain its proper function, but it cannot synthesize *de novo* or does so in insufficient quantities. For this reason, they are considered essential nutrients that must be acquired through the diet.

In most cases, vitamins are not single molecules but rather a group of related molecules called vitamers. There are 13 different vitamins: A, B₁, B₂, B₃, B₅, B₆, B₇, B₉, B₁₂, C, D, E and K, each with a distinct biochemical role in the cell (6). These roles typically range from gene regulation (vitamin A) to hormone-like function (vitamin D) to enzyme cofactors,

which is the case for all B vitamins. Although not chemically related, these are either a cofactor or precursor to one.

One of the essential B vitamins is vitamin B₆. Vitamin B₆ is a collective term used to refer to several vitamers (**Figure 1**), namely the pyridoxine (PN, alcoholic form), the pyridoxal (PL, aldehydic form), and the pyridoxamine (PM, amino form), as well as, their respective phosphate esters, pyridoxine 5'-phosphate (PNP), pyridoxal 5'-phosphate (PLP), and pyridoxamine 5'-phosphate (PMP). All these forms of vitamin B₆ can be interconverted under physiological conditions. Nevertheless, the biologically active form of vitamin B₆ is pyridoxal 5'-phosphate (PLP) (7).

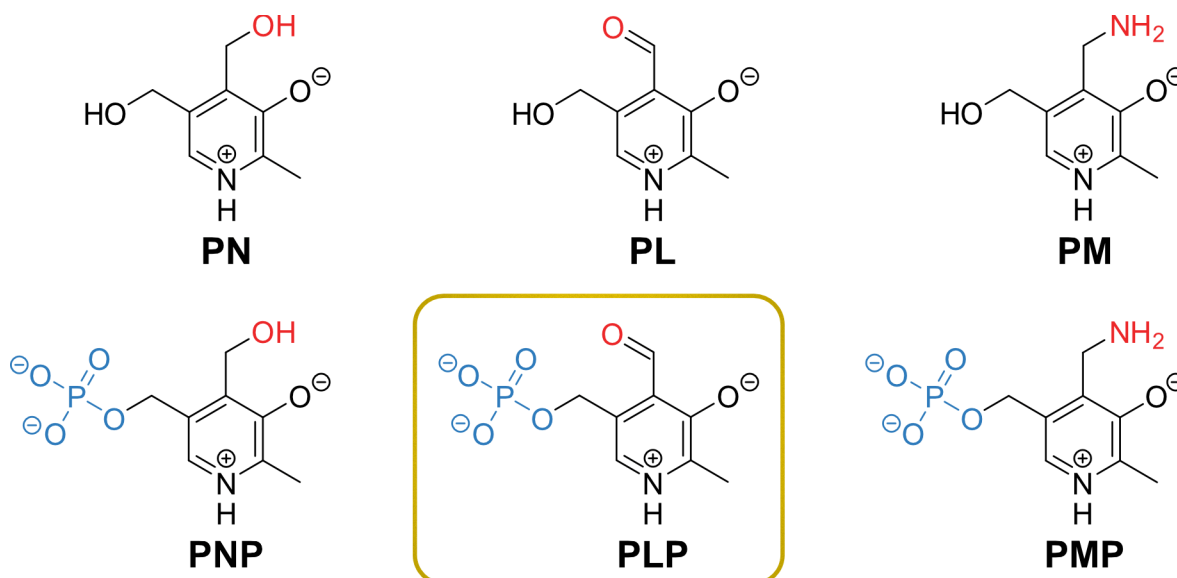


Figure 1. Structure of all B₆ vitamers, pyridoxine (PN), pyridoxal (PL), pyridoxamine (PM), pyridoxine 5'-phosphate (PNP), pyridoxal 5'-phosphate (PLP; the active form of vitamin B₆), and pyridoxamine 5'-phosphate (PMP).

PLP is a versatile cofactor required by a myriad of enzymes mostly related to amino acid metabolism, as we shall see in **section 2.3**. However, PLP is also involved in the biosynthesis of neurotransmitters, glucose metabolism, lipid metabolism, and haemoglobin formation (8). Moreover, it also plays a role in the immune system by acting as a cofactor for the enzyme L-histidine decarboxylase, which synthesizes histamine (9).

In the next section, we will focus on the *de novo* biosynthesis of vitamin B₆, which can occur by two distinct pathways found in different organisms. Moreover, the salvage pathway will also be briefly described.

2.2. Biosynthesis of PLP

There are three main pathways for the biosynthesis of vitamin B₆: two *de novo* pathways (**Figure 2**), the DXP-dependent pathway and the DXP-independent pathway, as well as the salvage pathway, the only one found in humans (7,8).

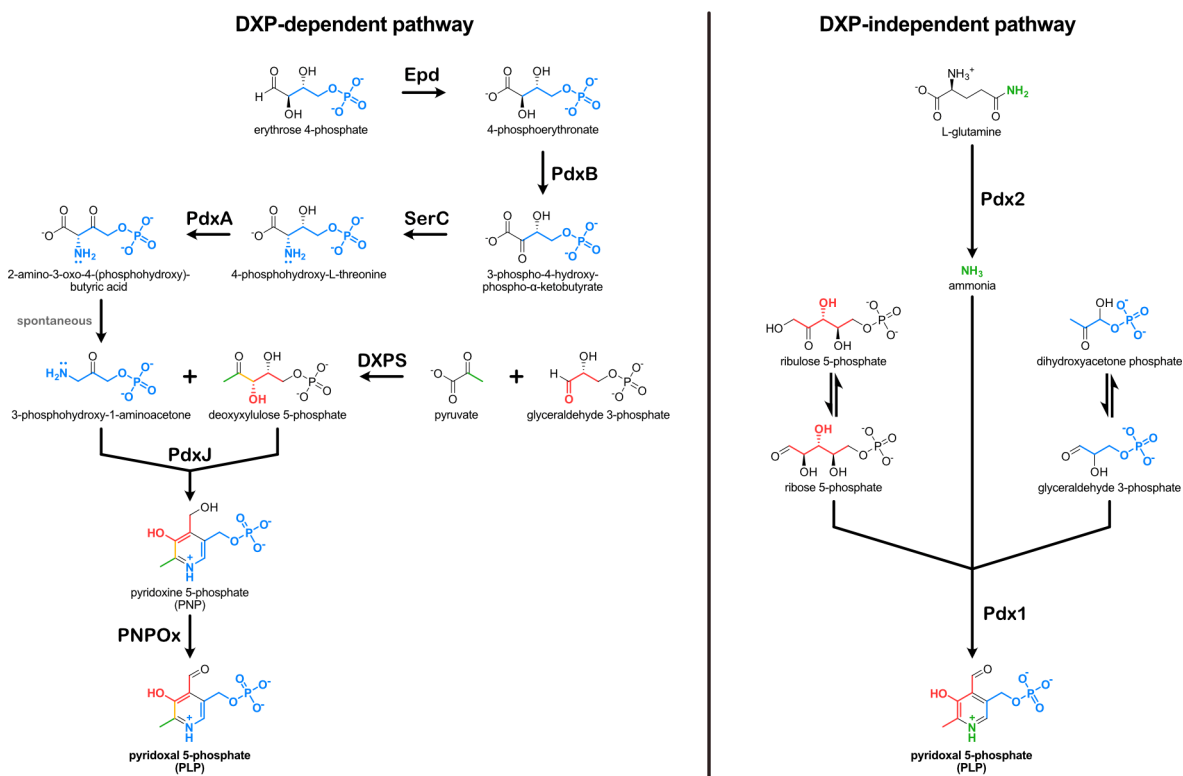


Figure 2. DXP-dependent and DXP-independent PLP biosynthetic pathways. Epd, erythrose 4-phosphate dehydrogenase; PdxB, 4-phosphoerythronate dehydrogenase; SerC, 3-phosphoserine aminotransferase; PdxA, 4-phosphohydroxy-L-threonine dehydrogenase; PdxJ, PNP synthase; DXPS, 1-deoxyxylulose 5-phosphate synthase; PNPOx, PNP oxidase; Pdx2 (glutaminase subunit) and Pdx1 (synthase subunit) of the PLP synthase complex.

Since humans lack the pathways needed to synthesize vitamins *de novo*, vitamin biosynthetic pathways are often regarded as potential targets for the development of new drugs against highly deadly diseases like Malaria and Tuberculosis (10–16).

In the following sections, we will examine these pathways, focusing on where they can be found, how they differ, and which enzymes are necessary for PLP synthesis. For the DXP-independent pathway, a more detailed description of each enzyme is also presented.

2.2.1. *De novo* pathways

2.2.1.1. DXP-dependent pathway

First of all, why is this pathway called DXP-dependent? DXP stands for deoxyxylulose 5-phosphate, one of the intermediates in the PLP biosynthetic pathway first found in *Escherichia coli*. This was the first pathway to be discovered, and for some time, it was thought to be the only one, ubiquitous in all organisms able to synthesize PLP (17). Later, however, it was found that the DXP-dependent pathway is predominantly only found in Gammaproteobacteria (18), such as *E. coli* and *Salmonella enterica*.

This pathway requires several steps that, through the convergence of two different branches, result in PLP. One of the branches is responsible for converting erythrose 4-phosphate into 3-phosphohydroxy-1-aminoacetone through the sequential action of the enzymes GapA, PdxB, PdxF/SerC, and PdxA (19–22). The other branch is responsible for the synthesis of deoxyxylulose 5-phosphate (DXP) through the condensation of D-glyceraldehyde 3-phosphate and pyruvate by the enzyme DXP synthase (DXPS) (22,23). The two resultant molecules, 3-hydroxy-1-aminoacetone phosphate and DXP, are then condensed into a pyridoxine 5'-phosphate (PNP) molecule by the enzyme PdxJ (24,25). Finally, PNP is oxidized into PLP by the enzyme PdxH (also known as PNPOx) (26). So, in sum, this pathway requires seven steps and seven different enzymes to catalyze the synthesis of one molecule of PLP.

2.2.1.2. DXP-independent pathway

The DXP-independent pathway, contrary to what was initially thought, is considered the “universal” pathway, found in most forms of life, like plants, fungi, protozoans such as the Apicomplexan parasites, *Plasmodium falciparum* and *Toxoplasma gondii*, and in most bacteria (except those of the Gammaproteobacteria class, as previously mentioned) (17,27,28). Notably, however, humans, much like all other mammals, are incapable of synthesizing PLP *de novo*, resorting to acquiring it from their diet in the form of B₆ vitamers. These vitamers are then converted into PLP by the salvage pathway, which will be briefly addressed in **section 2.2.2**.

Unlike in the DXP-dependent pathway, which consists of seven different enzymes, in the DXP-independent pathway, only one multifunctional enzyme is required, the PLP synthase (EC 4.3.3.6), a class-I (also known as trpG-type or triad) glutamine amidotransferase (GATase). Class-I GATases are heteromeric enzymatic complexes

consisting of two different types of subunits: the glutaminase and the synthase subunits. In PLP synthase, these subunits are commonly named Pdx2 and Pdx1, respectively (29).

Instead of erythrose 4-phosphate or DXP, PLP synthase requires three other molecules: glutamine, which is the nitrogen source for the pyridine ring of PLP; the pentose, ribose 5-phosphate (or its isomer, ribulose 5-phosphate), and the triose, glyceraldehyde 3-phosphate (or its isomer dihydroxyacetone phosphate). The fact that PLP synthase accepts both pentose phosphate isomers, as well as both triose phosphate isomers, clearly suggests that it has pentose and triose isomerization capabilities (30,31).

Regarding its 3D structure, PLP synthase is a 24-meric macromolecular complex composed of 12 Pdx1 and 12 Pdx2 subunits. Pdx1 subunits assemble as a dodecamer with a cylindric shape, built from two opposing hexameric rings. In addition, a Pdx2 subunit interacts with each Pdx1 subunit, giving PLP synthase a cogwheel-like shape (**Figure 3**) (17,32,33).

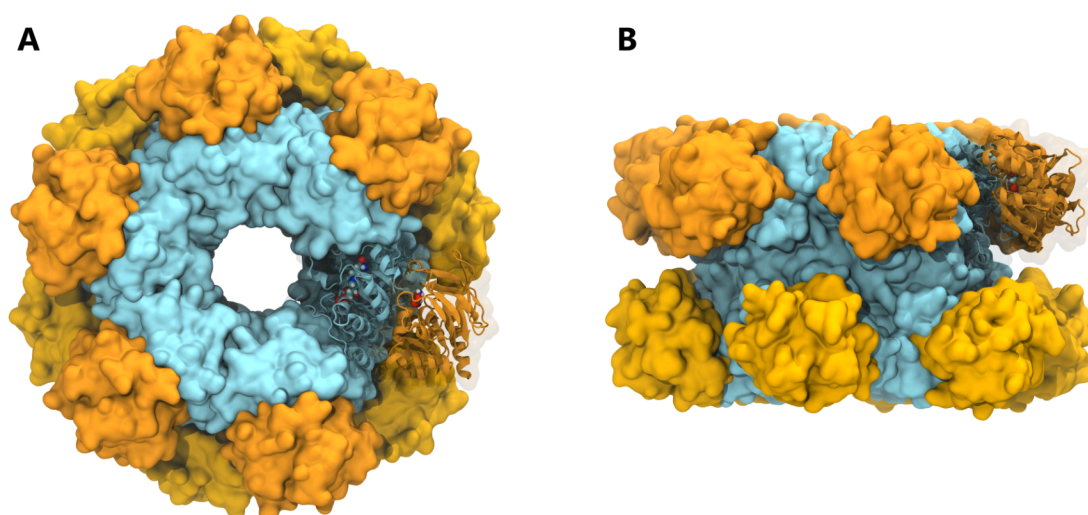


Figure 3. 3D representation of the PLP synthase macro-complex. (A) Top view. (B) Side view. The Pdx1 synthase subunits (represented in blue) assemble in a dodecameric form, two hexamers on top of each other, with one Pdx2 glutaminase subunit (represented in orange) bounded to each of them.

In the following sections, the role of Pdx1 and Pdx2 subunits in PLP biosynthesis will be described in detail, also addressing broad aspects like complex formation and the existence of a putative tunnel linking both subunits.

a. Pdx2

Pdx2, also known as PdxT, YaaE, and SNO1 (depending on the organism), is the glutaminase subunit of PLP synthase, responsible for the hydrolysis of glutamine into glutamate and ammonia (27,34–36). The ammonia produced by Pdx2 is then channelled to the active site of Pdx1 via a putative tunnel, where it is used as the nitrogen source for the pyridine ring of PLP (33,37–39).

However, the glutaminase activity of Pdx2 is only observed when it is interacting with Pdx1. This interaction is crucial for the stabilization of the Pdx2 active site, allowing for glutamine binding and hydrolysis (40). Moreover, according to Semmelmann et al., the activation of the glutaminase depends not only on the assembly of the PLP synthase complex but also on the presence of substrate (i.e., ribose 5-phosphate or ribulose 5-phosphate) at the Pdx1 synthase subunit (41). This tight regulation between Pdx2 and Pdx1 allows for ammonia production only when the Pdx1 active site is ready to receive it. Furthermore, by only producing ammonia when necessary and channelling it directly to the Pdx1 active site, the PLP synthase shields this highly toxic molecule from wreaking havoc within the cell.

Pdx2's active site is located at its interface with the N-terminus of Pdx1. Like all other class-I GATases, the conserved Cys₈₇-His₁₉₆-Glu₁₉₈ catalytic triad (numeration from the *Plasmodium falciparum* PLP synthase) constitutes the key element for the catalysis (40). Nevertheless, without the oxyanion hole, the reaction would not be possible. The oxyanion hole is a region of the Pdx2 active site composed of the nitrogen atom of the peptide bonds between the catalytic Cys₈₇ and the adjacent Ala₈₈, as well as between Gly₅₁ and Gly₅₂ (40). Its role is to stabilize the charge of the glutamine carbonyl group during catalysis and is one of the Pdx2 regions involved in its control, as the loop containing Gly₅₁ and Gly₅₂ only frees the active site when the PLP synthase complex is assembled (28).

Regarding the kinetics of the Pdx2 glutaminase activity, its turnover number (k_{cat}) varies drastically from one organism to another, ranging from 0.08 s⁻¹ (pH 8.0, 30 °C), for *Plasmodium berghei* (28), to 0.20 s⁻¹ (pH 8.0, 30 °C), for *Saccharomyces cerevisiae* (36). In between, there is *Plasmodium falciparum* with 0.11 s⁻¹ (pH 8.0, 30-37 °C) (40,42) and *Bacillus subtilis* with 0.13 s⁻¹ (pH 8.0, 37 °C) (33). Moreover, mutagenic studies have shown that when the catalytic His is mutated to an Asn, Pdx2 is not able to hydrolyze glutamine (33).

Finally, as part of this PhD project, the catalytic mechanism of Pdx2 from *Plasmodium falciparum* was studied in more detail (**Chapter C – Section 1.1 – Page 107**).

Catalytic triads: the role of cysteine

A catalytic triad consists of a set of three amino acid residues found in the active site of some enzymes that act in coordination to perform the enzymatic catalysis (43). This is necessary because no residue alone is a strong enough nucleophile at physiological pH (44).

Throughout this section, the term “catalytic triad” refers specifically to those that utilize either serine (Ser) or cysteine (Cys) as their nucleophile and histidine (His) as their base, as these are the most relevant to this thesis.

Each of the three residues has a specific role in catalysis. Of the three, only two residues are essential, the nucleophile and the base, reason why we can also find catalytic dyads, like the Cys-His dyad from the SARS-CoV-2 Main Protease (M^{pro}) (45,46). The third residue, usually an acidic residue like Asp or Glu, typically does not take an active role in the catalysis, mainly keeping the His imidazole ring in the correct orientation and helping to stabilize the positive charge that forms on His during the catalysis (47–49).

Typically, the residue functioning as a nucleophile is either a serine (Ser) or a cysteine (Cys) residue, while a histidine (His) residue works as a base.

Catalytic triads are primarily found in proteases, like trypsin (a Ser protease) or papain (a Cys protease), but are also present in other enzymes of the hydrolase class (EC 3), mainly those acting on carbon-nitrogen bonds, other than peptide bonds (EC 3.5), like the glutaminases.

At first glance, the mechanism employed by the catalytic triads Ser-His-Asp/Glu and Cys-His-Asp/Glu is incredibly similar (**Figure 4**) (50). In both cases, the nucleophile, which is deprotonated by the His residue, performs a nucleophilic attack on the carbon of the carbonyl adjacent to the bond that will be cleaved (C(O)–NH). This leads to the formation of a tetrahedral reaction intermediate, in which the substrate becomes covalently bonded to the enzyme (i.e., an adduct). In it, the electron density of the carbonyl double bond becomes concentrated on the oxygen atom, which is stabilized by the oxyanion hole region, as mentioned in the case of the Pdx2. Then, the positively charged His protonates the nitrogen atom of the leaving group (e.g., ammonia in the case of glutaminases or the rest of a peptide in the case of proteases), which can be concomitant or not with its

release and formation of an acyl-intermediate. After the departure of the amine product, there is enough space for water molecules to enter the active site. One of those water molecules will then perform a nucleophilic attack to the same carbon as before, at the same time as the catalytic His removes one of its protons, effectively transforming the water molecule into a hydroxide, which is a much better nucleophile. This leads to the formation of a second, unstable, tetrahedral reaction intermediate that quickly collapses into the product of the reaction.

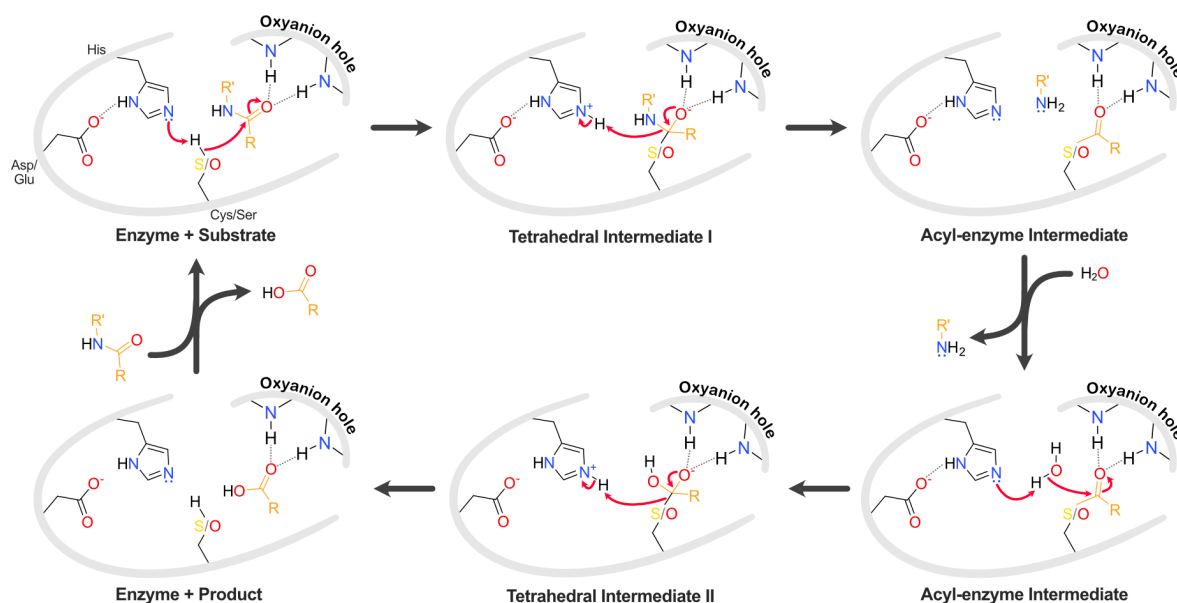


Figure 4. General catalytic mechanism employed by enzymes with Ser-His-Asp/Glu or Cys-His-Asp/Glu catalytic triads. At first glance, independently of the nucleophile used by the enzyme, the mechanism is very similar. However, there are some nuances, as we shall see below, especially regarding the activation of the nucleophile. Adapted from (50).

As you may have noticed, I did not explain how the nucleophile's deprotonation occurs, as this is the main difference between triads employing either a Ser or a Cys. In the active site microenvironment, the pKa of a catalytic Cys thiol group can be as low as ~3.3 (51). This effectively means that the catalytic Cys exists in the active site in its thiolate form (S^-) as an ion pair with the positively charged His residue (48,50). This Cys^-His^+ ion-pair state has been observed in enzymes like the M^{pro} (52) and the papain (53,54). On the other hand, the pKa for the hydroxyl group of the Ser sidechain is 15.9 (50), being virtually impossible for Ser to exist in its alkoxide form.

This critical difference means that enzymes employing a Ser-His-Asp/Glu catalytic triad have to follow a concerted mechanism, while enzymes employing a Cys-His-Asp/Glu

catalytic triad follow a stepwise one, first deprotonating Cys's thiol group and only then performing the nucleophilic attack (50). On the one hand, this means that by forming a stable ion-pair intermediate, Cys-His-Asp/Glu enzymes can indefinitely remain in this step while waiting for the substrate, avoiding the redundant protonation and deprotonation step each time a new substrate enters the active site. However, on the other, this also comes with a disadvantage – the enzyme's active site is more prone to oxidation. Therefore, the concerted mechanism of Ser-His-Asp/Glu enzymes may be regarded as an evolutionary advantage, being less prone to oxidation, especially in aerobic environments (50).

b. Pdx1

Pdx1, also known as PdxS, YaaD, SNZ1, SOR1, or PYROA, depending on the organism, is the synthase subunit of PLP synthase (35,36,55–57). With just three substrates, ribose 5-phosphate, glyceraldehyde 3-phosphate, and ammonia, which come from the glutaminase subunit, Pdx1 can create a single molecule of PLP. To achieve this prowess, Pdx1 has to carry out pentose and triose isomerization, imine formation, ammonia addition, aldol-type condensation, cyclization, and aromatization, all in one domain, without the requirement of any cofactor (58).

Pdx1 is known for having two distinct active sites, commonly named P1 and P2, due to their phosphate-binding nature (33,37). Because P1 and P2 bind ribose 5-phosphate (R5P) and PLP, respectively, they are also known as the substrate- and the product-binding sites. The distance between these active sites is approximately 20 Å, measured from where the phosphate groups of R5P and PLP are anchored (59). This distance always led to the question of how the intermediates are transferred between both sites, a question that only started to be answered five to six years ago. From what is currently known, the catalytic process begins with the covalent binding of R5P to Pdx1 at its P1 active site via a conserved Lys residue, resulting in an imine linkage. As mentioned in the Pdx2 section, this triggers the catalysis of the glutaminase subunit, ultimately sending an ammonia molecule down the transient tunnel directly to the P1 active site. There, it will be incorporated into the Pdx1-R5P structure, forming the Pdx1-pre-I₃₂₀ intermediate. This structure was only seen when the conserved Lys residue at the P2 active site was mutated by an Arg, which completely abolished PLP formation. However, in a wild-type Pdx1, this Lys, initially pointing to the P2 site, “swings” towards the P1 active site leading to the imine linkage between its ε-amino group and the five-carbon structure derived from R5P. This results in Pdx1-I₃₂₀, a chromophoric intermediary bridged between two Lys residues. Its formation requires conformational changes, freeing the P1 phosphate-binding

site and creating space for G3P to bind. Then G3P covalently binds to the nitrogen atom originally belonging to the ammonia molecule, forming the Pdx1-I₃₂₀-G3P intermediate. With this, the enzyme has all the elements that are required to build PLP, and all its left is to orient them correctly for ring cyclization to be possible. Moreover, this is where the P2 active site comes into play. After the Pdx1-I₃₂₀-G3P intermediate is formed, the Lys from the P1 active site releases the intermediate from its end, allowing the P2 Lys to swing once again to its active site where the final steps of PLP formation will take place (59).

All this incredibly complicated mechanism was the strategy that Nature found to use a single enzyme to catalyze a process so complex as the assembly of PLP. In contrast, the DXP-dependent pathway of Gammaproteobacteria requires six enzymes to achieve the same goal.

2.2.2. Salvage pathway

Humans, similarly to other mammals, are only able to obtain PLP from the B₆ vitamers acquired from their diet. Nevertheless, regardless of whether they are able to synthesize PLP *de novo* or not, most organisms possess a “salvage pathway” (8).

The salvage pathway mainly consists of the pyridoxal kinase (PLK) and the pyridoxine/pyridoxamine 5'-phosphate oxidase (PNPOx) that are responsible for the recycling of the B₆ vitamers (**Figure 5**).

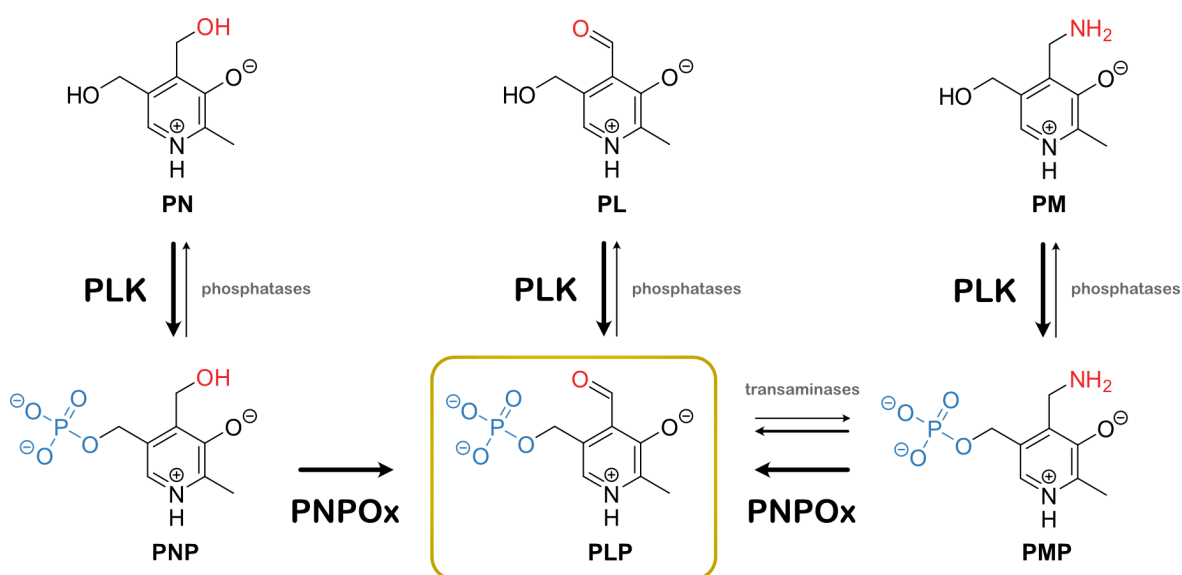


Figure 5. Biosynthetic salvage pathway for the “recycling” of B₆ vitamers into PLP. PLK, pyridoxal kinase; PNPOx, pyridoxine/pyridoxamine 5'-phosphate oxidase.

PLK is an ATP-dependent enzyme whose role is to phosphorylate the alcohol group located at the 5' position of PN, PL, and PM to form PNP, PLP, and PMP, respectively (60). PNPOx, on the other hand, is a flavin mononucleotide (FMN)-dependent enzyme that oxidizes either the 4'-hydroxyl group of PNP or the 4'-primary amine group of PMP to the aldehyde present in PLP. Moreover, besides being part of the salvage pathway, PNPOx also catalyzes the last step in the DXP-dependent PLP biosynthesis (26).

2.3. PLP-dependent enzymes

Pyridoxal-5'-phosphate (PLP)-dependent enzymes currently account for 233 different catalytic activities with an Enzyme Commission (EC) number, according to the BRENDA database (61). As of 5th December 2022, the database contains 8331 discovered enzymes, which means that almost 3% of all known enzymes require PLP as a cofactor. Furthermore, PLP-dependent enzymes can be found in five of the seven classes categorized by the Enzyme Commission (**Figure 6**), with only no ligases (EC 6) or translocases (EC 7) reported up to this date, thus representing a notable example of divergent evolution.

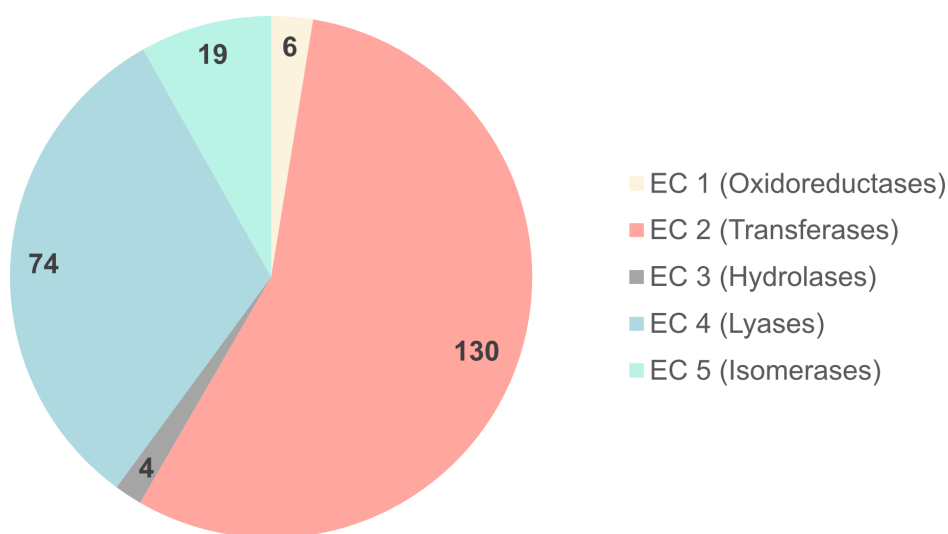


Figure 6. Distribution of all known PLP-dependent enzymes with an EC number attributed across the different enzyme groups.

One of the most used tools for monitoring the progress of the catalytic mechanism of the different PLP-dependent enzymes is UV-visible spectroscopy. The PLP cofactor absorbs light in the visible region of the electromagnetic spectrum, causing a colour change that can be used to study the progress of the reaction. Since the number of conjugated double bonds (i.e., every other bond is a double bond) in a molecular system is related to the

wavelength at which it absorbs light, alterations to the PLP's conjugated pi-bonding system will alter its absorption profile. During the catalytic cycle, electrons from the substrate are highly delocalized through the conjugated system of PLP, which is one of its essential roles as a cofactor, as we will see. This electron delocalization leads to an increase in the wavelength at which the cofactor absorbs light. Thus, by tracking changes in the cofactor's absorbance profile, one can gain insight into the catalytic mechanism of PLP-dependent enzymes. Furthermore, by analyzing the absorbance changes throughout the reaction, intermediates in the catalytic mechanism can be identified. Absorbance changes can indicate when intermediates form and when they react, allowing us to gain a better understanding of the sequence of events in the catalytic mechanism, as demonstrated by the example in **Figure 7**.

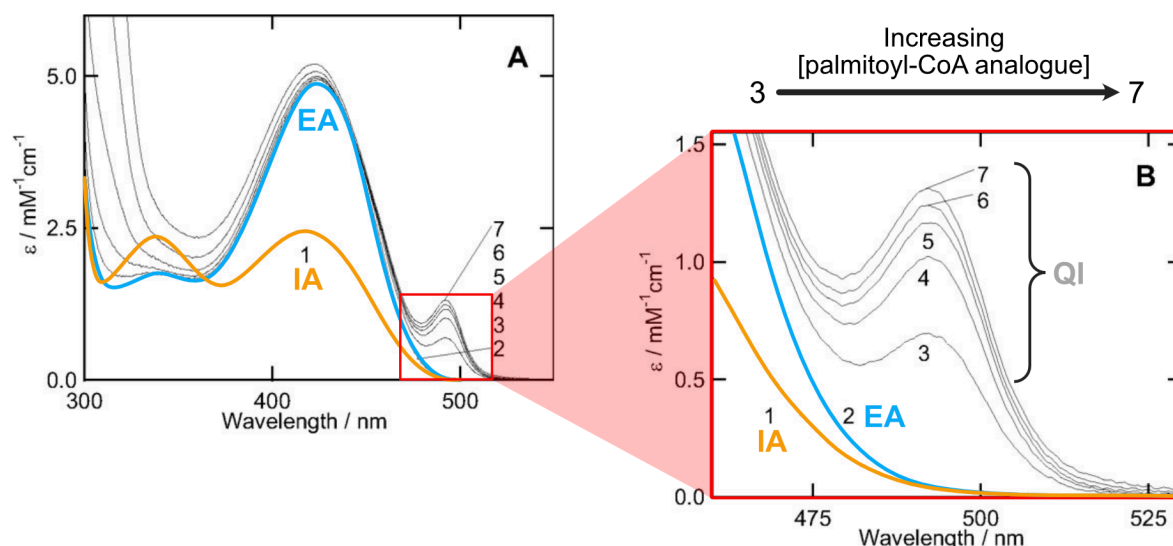


Figure 7. Absorption spectra of the PLP-dependent Serine palmitoyltransferase (SPT), a key enzyme in sphingolipid biosynthesis, adapted from the work of Ikushiro et al. (62). SPT catalyzes the decarboxylative condensation of L-serine with palmitoyl-CoA to generate 3-ketodihydrosphingosine. First, L-serine reacts with the internal aldimine (IA), forming the external aldimine (EA), then only if palmitoyl-CoA is present does the reaction continue with the formation of the quinonoid intermediate (QI), and ultimately the product. This was discovered by UV-visible spectroscopy using an analogue to palmitoyl-CoA that is not able to react with L-serine but whose presence triggers the formation of the QI, as observed in this absorption spectra. Furthermore, increasing concentrations of the analogue lead to a higher absorption at the typical QI wavelength, indicative of a higher QI formation.

Another aspect that makes PLP such a good cofactor is its very reactive aldehyde group that readily reacts with primary and secondary amines. This aspect allows PLP to easily react with amino acids aiding in the catalysis of a considerable number of different

reactions as an enzymatic cofactor. Nevertheless, this also has a dark side to it. Due to the high reactivity of PLP's 4'-aldehyde moiety, it easily forms covalent bonds with other compounds or proteins, having been classified as one of the 30 most damage-prone metabolites (63). For instance, PLP was shown to inhibit enzymes involved in DNA replication, like the eukaryotic DNA polymerases alpha and epsilon (64) and the topoisomerase IB (65). For all these reasons, the levels of free PLP within a cell must be kept low, as it would otherwise disrupt cellular homeostasis, leading to toxicity (66). In fact, microorganisms that synthesize PLP *de novo* do it close to the minimal necessary rate, maintaining its cellular concentration low (67).

2.3.1. *Versatility of the enzymes that employ PLP as a cofactor*

PLP-dependent enzymes are known for their ability to catalyze several different types of reactions (**Figure 8**), such as decarboxylation, racemization, transamination, β -, and γ -elimination/replacement, and retro-aldol cleavage, among others (68).

The Dunathan hypothesis suggests that the covalent bond cleaved in PLP-dependent enzymes should be perpendicular to the PLP plane (69,70). When the target σ bond (C_α – C_β , C_α – C_{COO^-} , or C_α – H_α) is in this orientation, it occupies the same region of the π -bonds that make up the PLP cofactor's conjugated system. Consequently, the electron density of the σ bond can be highly delocalized throughout the PLP cofactor, weakening the covalent bond and making the reaction faster and more favourable thermodynamically. Thus, the correct orientation of the target atom/group is vital for the reaction to occur. This is linked to the interactions established between the external aldimine and key amino acid residues present in the active site of the different PLP-dependent enzymes.

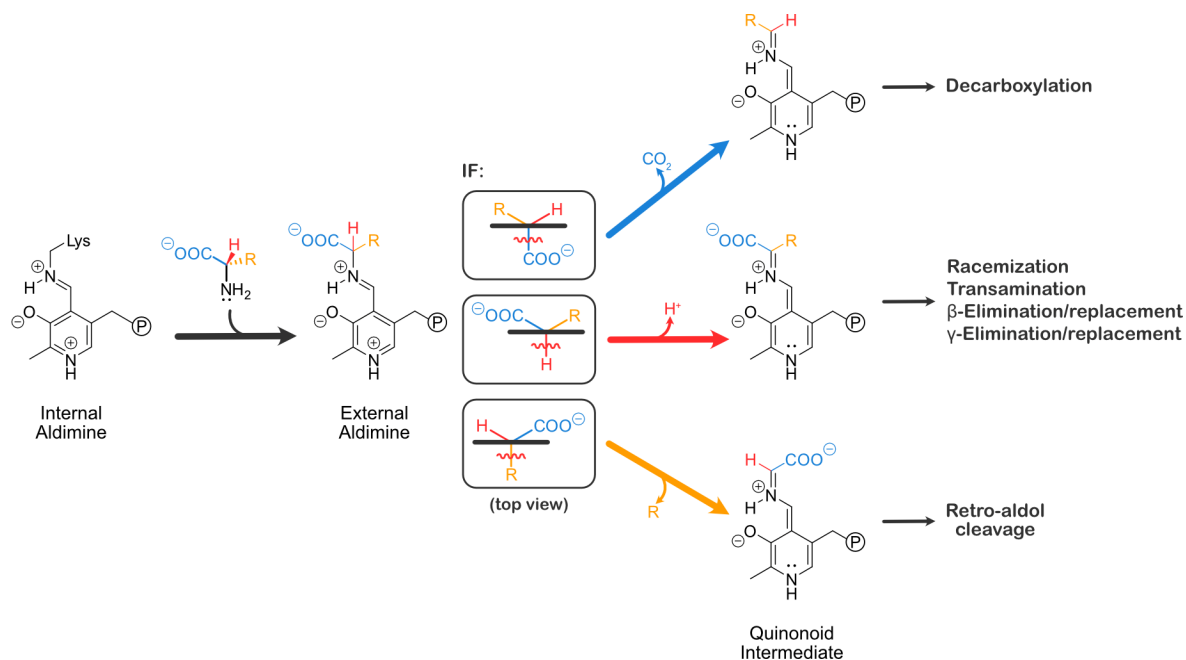


Figure 8. Overview of the different types of reactions catalyzed by PLP-dependent enzymes. According to the Dunathan hypothesis, the bond to be cleaved is the one perpendicular to the plane of the pyridinium ring of PLP. Depending on the enzyme, this leads to the formation of one of three types of quinonoid intermediates. These intermediates can then undergo several different reactions, demonstrating the high versatility of PLP-dependent enzymes.

For example, the decarboxylation reaction results from the cleavage of the bond between the substrate's α - and carboxyl carbon atoms. As a result, one carbon dioxide molecule is released, and the resulting intermediate is protonated in the α -carbon (9).

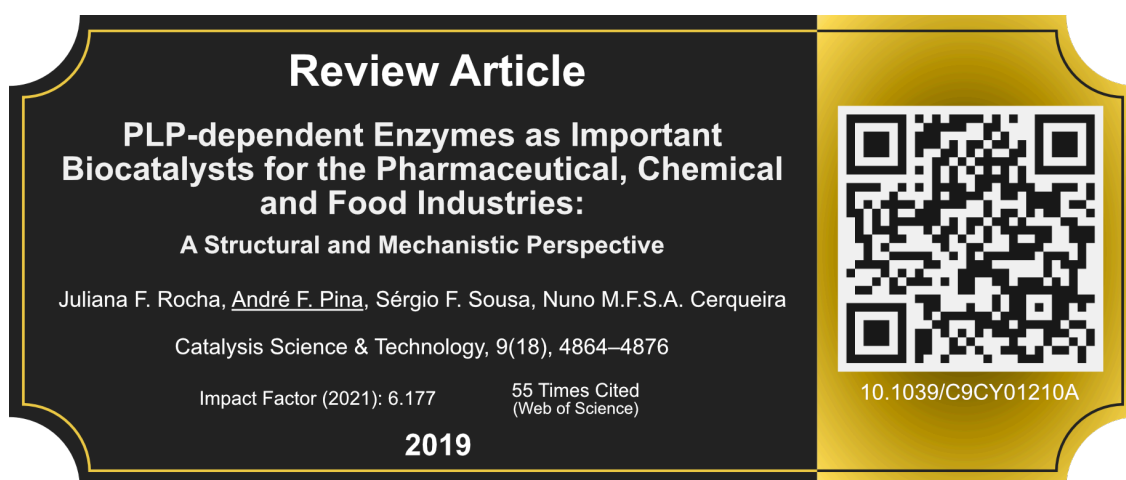
In racemization, the chiral centre of the amino acid is inverted from L to D or vice versa. Enzymes accomplish this by removing, from one side of the molecule, the proton bonded to the α -carbon, followed by re-protonation from the opposite side.

PLP-dependent enzymes can also catalyze elimination or substitution reactions at the α , β , and γ positions of an amino acids' sidechain. However, there is no single mechanism for this kind of reaction, varying from enzyme to enzyme and from substrate to substrate. In this PhD project, the catalytic mechanism of a β -elimination reaction was studied in detail (**Chapter C – Section 2.1 – Page 131**). This reaction is catalyzed by XometC, a promiscuous PLP-dependent enzyme from a pathogen responsible for a devastating disease affecting rice crops worldwide (71,72).

Finally, we have the transamination reactions, which can be divided into two different half-reactions. In the first half, the amine group of an amine donor (e.g., amino acid) replaces the lysine residue covalently bonded to PLP (i.e., internal aldimine), which culminates in

the generation of pyridoxamine 5'-phosphate (PMP) and the corresponding ketone. In the second half, the amine group now in PMP is transferred to an amine acceptor (e.g., ketone, keto acid, or aldehyde), regenerating the internal aldimine and forming the corresponding amine, completing the catalytic cycle (73).

Moreover, a review regarding examples of the catalytic power of PLP-dependent enzymes for industrial purposes was prepared in this thesis (**Appendix:** page 273). This review focuses on several PLP-dependent enzymes with considerable potential as biocatalysts actively used by the pharmaceutical, agricultural, and food industries, some of which are great examples of enzymes that catalyze the activities mentioned above.



2.3.2. Common aspects of the catalytic process of PLP-dependent enzymes

Although PLP-dependent enzymes can catalyze a panoply of different reactions, the majority share a common mechanism of activation: i) the covalent bonding of PLP to the enzyme by a conserved Lys residue (i.e., enzyme's activation) and ii) the subsequent bonding of the substrate to PLP.

The first stage, the enzyme's activation, has already been computationally studied by E.F. Oliveira et al. and consists of three sequential and reversible steps (**Figure 9 – A-C**) (74). First, as the PLP cofactor enters the active site, the conserved Lys residue performs a quick nucleophilic attack ($\Delta G^\ddagger = 0.2 \text{ kcal.mol}^{-1}$) to the reactive carbonyl group of PLP (**Figure 9 – A**). Then, a proton is transferred from the amino group of the Lys to the hydroxyl group of PLP (**Figure 9 – B**). This transference is typically mediated by a water molecule, albeit some PLP-dependent enzymes use other species, namely a Cys residue. Finally, the internal aldimine (IA) is formed with the release of a water molecule (**Figure 9 – C**). This new structure, anchored to the enzyme's active site, then waits for a substrate

molecule to initiate a catalytic cycle. However, the entire process mentioned so far is reversible, meaning that the enzyme can follow the opposite mechanism and release the PLP from its active site (**Figure 9 – D-F**). Nevertheless, the release of PLP from the enzyme is thermodynamically and kinetically less favourable than the formation of the internal aldimine, which requires a Gibbs free energy of activation of $13.5 \text{ kcal.mol}^{-1}$ and is exergonic in $9.1 \text{ kcal.mol}^{-1}$ (74).

After forming the internal aldimine (IA), a substrate molecule can enter the active site and ultimately replace the Lys residue, becoming covalently bonded to PLP, a structure known as external aldimine (EA) (**Figure 9 – G-I**). The mechanism leading to the formation of the EA was proposed to be uniform across all PLP-dependent enzymes (75–79).

According to the proposal from N.M.F.S.A. Cerqueira et al. (75), the formation of the EA starts with a nucleophilic attack of the amino group of the substrate to the carbon of the IA imine bond (**Figure 9 – H**). This results in a reaction intermediate known as gem-diamine, where both Lys and the substrate are covalently bonded to PLP.

Then, a proton is transferred from the amino group of the substrate to the amino group of the Lys residue (**Figure 9 – I**). This transfer has been proposed to occur either directly or mediated by a Tyr residue, a Cys residue, or a water molecule (75,76,79). N.M.F.S.A. Cerqueira et al. found out that the mediator that provides the most favourable pathway is the water molecule (75). This reaction leads to the Lys residue's release with the EA's formation, enabling the enzyme to catalyze its specific reaction.

However, the study of the catalytic mechanism of XometC presented in this thesis (**Chapter C – Section 2.1 – Page 131**), which includes all the steps from the external aldimine formation, suggests that the mechanism proposed above may not be the same in all PLP-dependent enzymes as we shall see. Since XometC's active site lacks any residue able to mediate the proton transference between the amino groups of the Lys and the substrate, we found that the hydroxyl group of the PLP ring can perform this mediation in a thermodynamically and kinetically favourable way. Nevertheless, XometC's mechanism still proceeds via the experimentally observed gem-diamine intermediate, as proposed above.

Lastly, after the enzyme finishes its specific set of reactions, the product, which is still covalently bonded to PLP, must be released from the enzyme, allowing a new catalytic cycle to begin. To do so, the Lys residue must attack the carbon of the imine bond, forming a new gem-diamine (**Figure 9 – K**). Then, after the proton transfer between the amino group of Lys and the product's amino group, the latter is released, restoring the

initial IA conformation (**Figure 9 – L**). Finally, the enzyme can undergo a new catalytic cycle if enough substrate is available. Otherwise, it returns to its apo form by releasing the PLP cofactor.

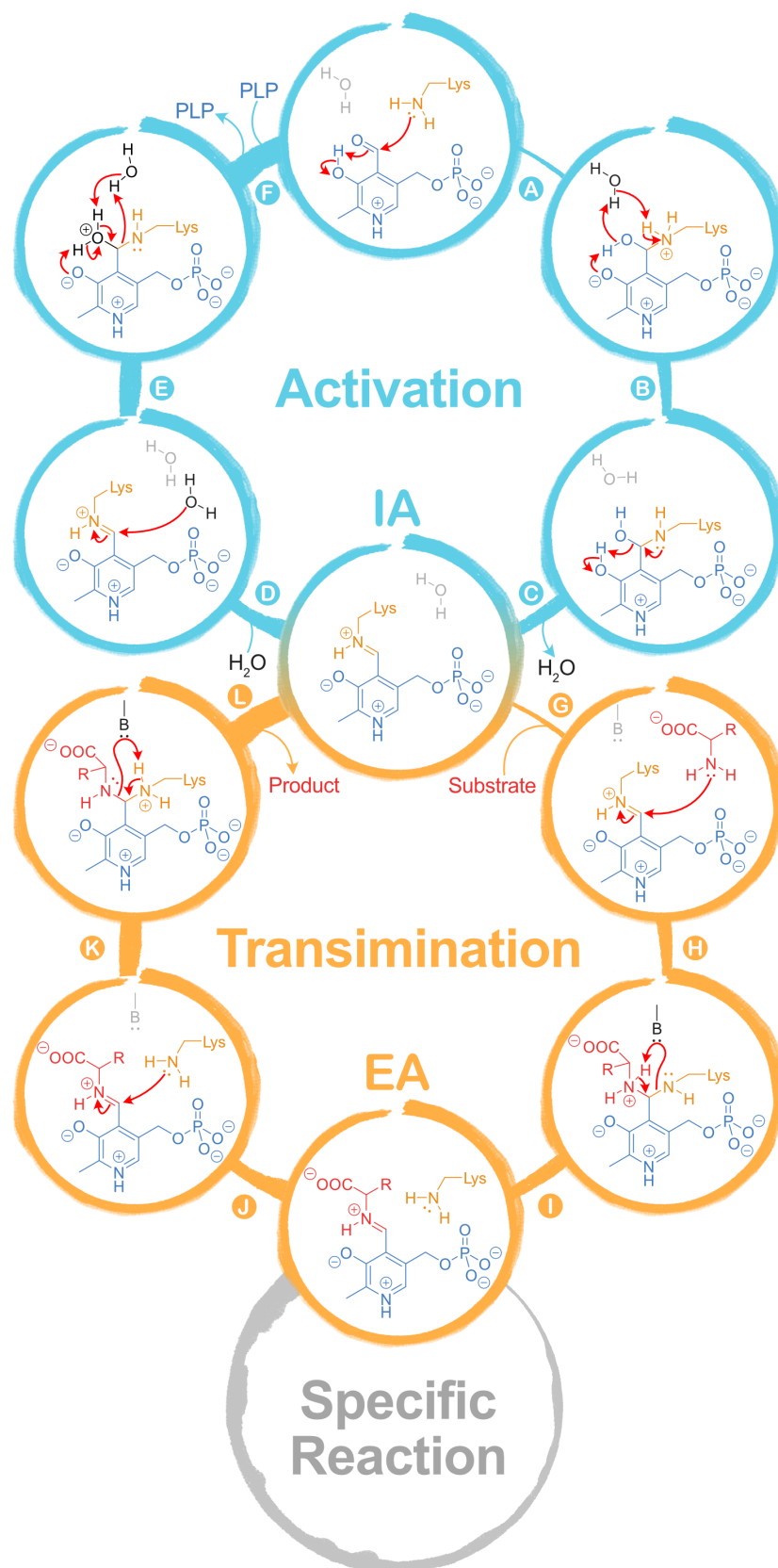


Figure 9. Schematic representation of the general mechanism for the activation of a PLP-dependent enzyme with the formation of the internal (IA) and its transamination with a substrate, forming the external (EA) aldimines.

3. DNA

Deoxyribonucleic acid (DNA) is a nucleic acid that contains genetic instructions used in the development and functioning of all known living organisms. In its most common form, DNA is made up of two strands of nucleotides coiled around each other forming a double helix. Each strand is made up of four nucleotide bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The sequence of these bases along the strands is the basis for the genetic information, with each triplet of nucleotides coding for a particular amino acid.

Nucleotides are the basic building blocks of DNA. Each nucleotide consists of a sugar molecule, a nucleobase, and a phosphate group (**Figure 10**). The sugar molecule is a five-carbon sugar called deoxyribose, while the nucleobase is, as previously mentioned, either adenine, guanine, cytosine, or thymine. Finally, the phosphate group provides the linking element that connects the nucleobases, creating a long chain of nucleotides – a strand of DNA.

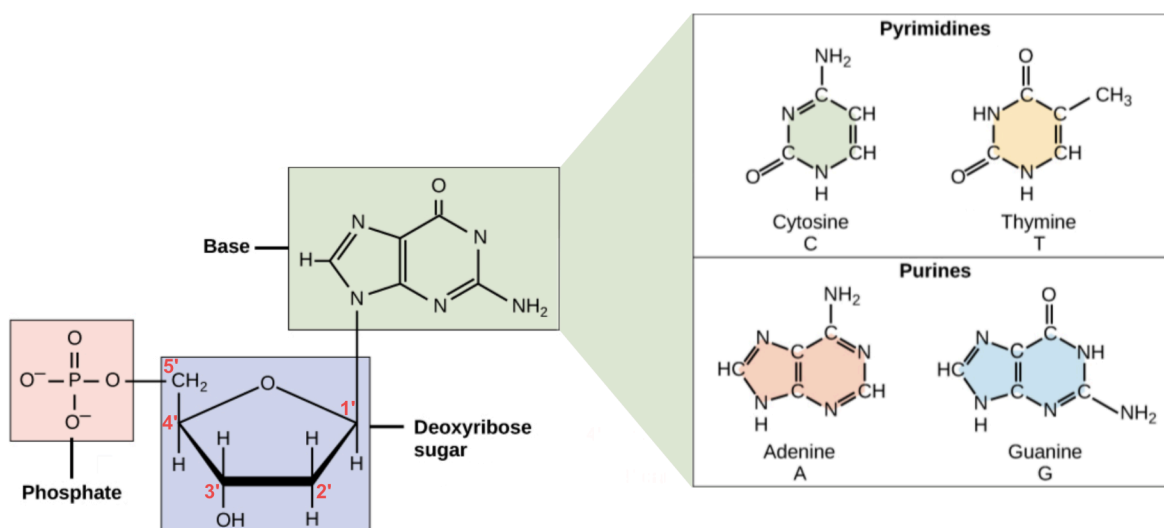


Figure 10. Nucleotides comprise a nucleobase, a pentose sugar (deoxyribose in the case of DNA), and a phosphate group. Carbon residues in the pentose are numbered 1' to 5'. New nucleotides are added at the 3' hydroxyl group by covalently bonding to the 5' phosphate group of the incoming nucleotide. Bases can be divided into two categories: purines and pyrimidines. Adapted from (80).

DNA strands have two ends, the 5' and 3'. The 5' end is known as such because the phosphate group of the terminal nucleotide at this end has a phosphate group attached to the 5' carbon of its sugar molecule. Similarly, the 3' end of the DNA strand is referred to as the "3' end" since the last nucleotide at this end has a hydroxyl group attached to the 3' carbon of its sugar molecule. The direction of the strand is determined by the orientation of the 5' and 3' ends.

The structure of DNA was first proposed in 1953 by James Watson and Francis Crick, who famously suggested the double helix structure (81). Their model was based on the X-ray crystallography of Rosalind Franklin and Raymond Gosling (82), which allowed them to determine the 3D structure of DNA. This discovery revolutionized the field of molecular biology, bringing genetic engineering and biotechnology into the public consciousness.

DNA can take on different conformations, including the canonical B-DNA (83), A-DNA (83), and Z-DNA (84). However, only B- and Z-DNA have been found under physiological conditions. B-DNA is the most common form of DNA, a right-handed double helix with 10.4 bases per turn and a phosphate-to-phosphate distance of approximately 7 Å, occurring at the high hydration levels of a cellular environment. A-DNA is a less common form, wider than B-DNA, typically found in partly dehydrated conditions. Finally, Z-DNA is the rarest form of DNA, typically found in regions of DNA that are actively transcribed. Contrarily to the other two conformations, Z-DNA has a left-handed double helix.

It is estimated that approximately 13% of the human genome can fold into non-B DNA structures, such as cruciforms, hairpins, G-quadruplexes, and Z-DNA, among others (85). Although these structures have been linked to important cellular processes, non-B DNA conformations have also been shown to hinder replication, resulting in errors and mutagenesis.

Moreover, DNA can also be found in a single-stranded form (ssDNA). ssDNA can be formed in various contexts, including during DNA replication, recombination, and transcription. ssDNA is more susceptible to enzymatic cleavage, thus being more susceptible to damage than dsDNA, but also making it easier to manipulate in genetic engineering. For this reason, ssDNA has been the basis of DNA aptamers, short oligonucleotides that can bind to specific targets with high affinity and selectivity.

As we have seen, DNA is a powerful and versatile molecule. As a result, the ability to study and manipulate DNA has become increasingly important in the modern world, allowing us to understand genetic diseases better and potentiating the development of medicines to treat them. By continuing to explore the potential of DNA, we are paving the way for a future of improved health and sustainability.

Although the primary goal of this PhD project was to study the PLP synthase macro-complex involved in the DXP-independent biosynthesis of PLP, the opportunity to study DNA-related systems presented itself. Therefore, two additional studies that resulted from external collaborations are presented in this thesis. The first pertains to the study of the 3D structure of non-B DNA conformations associated with mutagenic instability in the

mitochondrial genome (**Chapter C – Section 3.1 – Page 161**). The other consisted in the use of computational methodologies, namely molecular docking, molecular dynamics simulations, and binding free energy calculations (described in detail in the subsequent chapter), to guide the selection of a DNA aptamer with high affinity to cathepsin B (CatB), a highly expressed protein in cancers like the prostate cancer (**Chapter C – Section 4.1 – Page 185**). Therefore the following two sections provide a quick overview of non-B DNA conformations and how they may be related to mutagenic instability, as well as DNA aptamers and how computational methodologies can aid in their development.

3.1. Non-B DNA conformations: a cause of mutagenic instability

Certain DNA sequence motifs can fold into three-dimensional (3D) conformations that differ from the canonical right-handed B-DNA double helix. Those structures are known as non-B DNA conformations and consist of a few dozen to a few hundred nucleotides (85–87).

Several types of non-B DNA conformations have been identified (**Figure 11**). Their structure results from their nucleotide motif sequences and the solvent environment in which they are located. Non-B DNA conformations can take the form of simple conformations such as DNA hairpins or more complex structures such as G-quadruplexes (85–87).

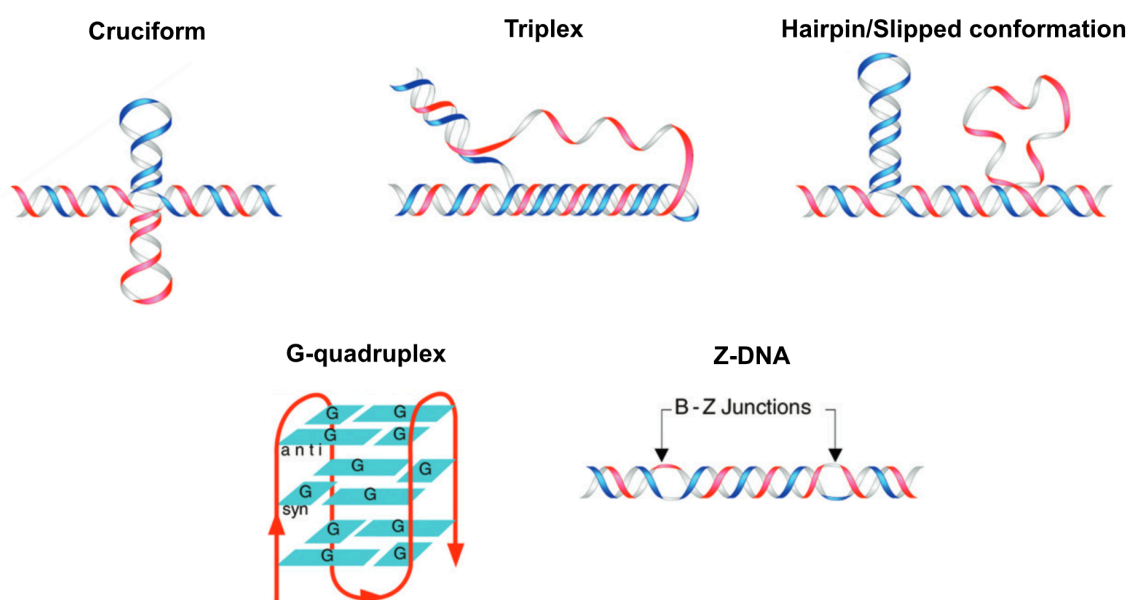


Figure 11. Schematic representation of several of the most common non-B DNA conformations. Adapted from (88).

These conformations do not follow a random distribution across the human nuclear genome, and several can have functional implications for cellular functions and human diseases. As previously mentioned, approximately 13% of the human nuclear genome can fold into non-B DNA conformations, affecting the process of DNA replication (85).

Non-B DNA conformations also occur in the human mitochondrial genome (89). Although it represents only a tiny part of the human genome, it contains 16 569 base pairs and encodes for two ribosomal RNAs (i.e., 12S and 16S), 22 transfer RNAs (tRNAs), and 13 proteins, all involved in the oxidative phosphorylation pathway. Unlike the linear chromosomes found in the nucleus, the mitochondrial DNA (mtDNA) occurs as a circular chromosome composed of two strands known as the H- and L-chains. At the end of the H-chain, there is a non-coding region known as the control region (90).

The control region is responsible for regulating the transcription and replication of mtDNA. Its mutation rate is also higher than in other parts of the mtDNA genome, which may be why it is the most variable region among different individuals (91). Moreover, the control region is also thought to play a role in the development of mitochondrial diseases, as mutations in this region can lead to an altered mtDNA replication rate and dysregulation of the mitochondrial network (92–94). Thus, understanding the structure and function of the control region is essential for the study of the mitochondrial genome and its associated diseases.

A bioinformatics analysis performed by Pereira et al. demonstrated a correlation between the occurrence of non-B DNA motifs in the mtDNA and the rate of nucleotide substitutions (89). In fact, most replication errors can be associated with mtDNA regions where non-B DNA conformations can be formed. During mtDNA replication, there is the occurrence of extensive single-stranded DNA (ssDNA) stretches, which are more prone to the formation of non-B DNA conformations (95). These conformations can then hinder the polymerase activity by being a physical obstacle to fork progression (95). One example of a non-B conformation that influences the mitochondrial replicative process is Structure A, a cruciform conformation formed in the control region of mtDNA (89), which will be the focus of **Chapter C – Section 3.1 – Page 161**.

In summary, non-B DNA conformations have been highly linked to events of mutagenic instability, both in nuclear and mitochondrial DNA, being associated with several genetic diseases that result from replication errors, repair errors, and genomic instability.

3.2. DNA aptamers: a promising therapeutic tool

Targeted therapeutic approaches, such as personalized medicines and gene therapies, have been gaining increasing attention as potential disease treatments. In this context, nucleic acids such as DNA and RNA have emerged as promising candidates for the development of novel therapies (96–101).

DNA aptamers are short oligonucleotides that can bind to specific targets with high affinity and selectivity. They are clinically relevant due to their ability to recognize and bind to a variety of molecules, including proteins, metabolites, and small molecules. This makes them an attractive option for medical applications, as they can target and modulate various biological processes. For example, DNA aptamers can target cancer cells, bind to receptors on the cell surface, and deliver therapeutic agents such as drugs or enzymes (102). DNA aptamers can also target specific immune cells in the body, such as macrophages, allowing for specific modulation of the immune system (103).

The most common method for generating and isolating DNA aptamers is through a process known as Systematic Evolution of Ligands by Exponential Enrichment (SELEX) (104–106). This process randomly generates a large pool of DNA sequences, which are then exposed to the target molecule. The DNA sequences that bind to the target are then selected and amplified, and the process is repeated until the final aptamers are obtained.

However, the SELEX process can be very expensive and time-consuming. For this reason, *in silico* tools like 3D structure prediction, molecular docking, molecular simulations, and binding free energy calculations have been gaining increasing interest in aptamer design as they can promptly provide valuable insights about the binding affinity of a given aptamer to the target of interest, all at a fraction of the cost. By combining *in silico* approaches with experimental methods like SELEX, the generation of aptamers can be accelerated and optimized, making them an even more attractive therapeutic tool (107,108).

Despite the advantages of DNA aptamers, their use as therapeutic agents is not without its challenges. Aptamers can be prone to enzymatic degradation, which reduces their stability and efficacy. Additionally, they can be rapidly excreted from the body through renal filtration (109).

Nevertheless, DNA aptamers represent a promising tool for clinical applications due to their high affinity and selectivity, low cost and fast generation time, and wide range of potential targets. Further research on their structure and function may lead to the

development of more stable and effective aptamers with promising therapeutic use, as seen in the collaborative work presented in **Chapter C – Section 4.1 – Page 185**.

B. METHODOLOGY

This page was intentionally left blank.

Science has experienced rapid growth since the beginning of the last century. We have looked where no one ever looked before, to the immeasurable depths of our Universe and to the infinitesimal scale of the quantum world, where subatomic particles abide by the laws of quantum mechanics.

At the same time, computers have surged to the point where everyone has one in the palm of their hand more powerful than the one that put men on the Moon. They are the cornerstone of our current world. Looking back only a few years, we can understand how computers have been getting faster and faster. CPUs (Central Processing Units) are getting better, with clock speeds surpassing 5 GHz and their core count reaching astounding numbers, with 6-core processors being the norm even on the slimmest laptop. When considering non-consumer CPUs used in HPC (High-Performance Computing) these values reach 96 cores per CPU.

However, a more significant development in computing is the creation of GPUs (Graphics Processing Units). The first commercial GPU was launched in 1995 by NVIDIA more than 25 years ago, and since then, GPUs have come a long way, becoming ever more powerful and efficient.

For this reason, several scientific programs have started to take advantage of GPUs to significantly increase their performance. Most of the best-known GPU-accelerated Molecular Dynamics software are AMBER (110), GROMACS (111), NAMD (112), and LAMMPS (113). The use of GPUs in Molecular Dynamics has allowed us to simulate hundreds of nanoseconds in a reasonable time, even on consumer-grade GPUs, which tend to have a better price-to-performance ratio than the high-end GPUs designed for these workflows. This is true not only for Molecular Dynamics software but also in Quantum Chemistry (although the parallelization is less efficient here due to the type of mathematics involved). The only drawback is that this software tends to be paid, challenging to implement, or sometimes both. For these reasons, they are still gaining some traction among the computational chemistry community.

One could say that we are on the verge of the golden age of computational chemistry with the steady development of quantum computing. It will allow us to tackle ever more difficult and complex questions more accurately and using fewer approximations.

So, in the following sections, I will briefly describe the workflow used in this work and provide some technical background about the methodologies used. Unless stated otherwise, this methodology was based on the following books (114–118).

1. Preparation of the model system

In order to computationally study a biomolecular system, we first need to have access to its tridimensional (3D) structure. This structure can either be a prediction, typical in the case of nucleic acid structures, or be experimentally determined. Several experimental methods have been employed to reveal structural information pertaining to a wide range of biomolecular systems. X-ray crystallography, which accounts for most known structures, nuclear magnetic resonance (NMR), and cryo-electron microscopy (cryo-EM), are the three main methods used.

Due to the high number of currently available structures, there was a need for a database of structural data of large biomolecular systems, such as proteins and nucleic acids. For this reason, the Protein Data Bank (PDB) was created in 1971 (119). As of 2 November 2022, the PDB comprised a total of 197 512 structures, from which 170 364 are X-ray structures, 13 833 NMR structures, and 12 990 cryo-EM structures. These structures submitted by scientists worldwide are freely available via the website <https://www.rcsb.org>.

A single protein can have multiple entries on the PDB. These can be from different organisms, have different resolutions, be complete or incomplete, have mutations, ligands, or be determined with a different method. So how can we choose the PDB structure that is best adequate for our purposes? As a rule of thumb, one should choose a structure that belongs to the organism of interest or a close relative, have a low resolution, preferably below 2 Å, is complete (i.e., does not have any missing regions), does not have mutations, or these are easily reversible and has the ligand of interest or one that closely resembles it. However, we are not always so lucky to find such a structure, and some accommodations must be made. For example, if the structure is incomplete, homology modelling can be performed to complete the system.

Most structures used in computational studies are convenient from X-ray crystallography as it allows for high-resolution structures (i.e., with a small resolution value). However, this method has a limitation. Most of its structures do not contain hydrogen atoms (from here on out, referred to as protons). For this reason, normally, these atoms must be added prior to the beginning of the study. Furthermore, the protein environment affects the protonation state of its residues. Therefore, it is crucial to have the correct protonation since it can compromise the work's accuracy. In the absence of experimental evidence, we must predict the protonation state of all system residues. This can be accomplished using software like PROPKA (120,121). Nevertheless, these predictions should always be

manually verified to ensure that the correct protonation state is attributed. In case of doubt, the different protonation states should be studied. These protons are then added using the LeaP tool of the AmberTools software package (110).

Lastly, since we are starting from an X-ray crystallography structure and ultimately want to model it in solution, it would be unwise just to start performing molecular dynamics simulations or other studies. The most stable conformation of the system in the crystal may not be the same in solution or in the physiological conditions of the cytosol. Therefore, we try to model it as close to the real world as possible. To do so, we add protons, a solvent, and ions, which alters the model relative to the initial X-ray structure. To make our final model ready for simulations, we have to optimize its structure, removing potential clashes that arise from these modifications. To do that, we will use molecular mechanics.

This page was intentionally left blank.

2. Molecular Mechanics

Many critical molecular processes in Nature involve chemical reactions, such as the formation or breakage of covalent bonds. However, traditional force fields, explained in detail in the following section, cannot explicitly model chemical reactions, being then typically studied with quantum mechanics (QM) methods. These methods consider, at least, the atoms' outer electron shell and the interactions they establish. However, QM methods present a problem: they quickly become prohibitively expensive with the increase of the number of atoms in the system. Therefore, systems studied with QM methods typically consist of only a few hundred atoms. Another option is using a reactive force field like ReaxFF (122), a molecular mechanics (MM)-based method that allows the study of bond formation/breaking. These methods are mostly used in the study of thermal (123) and oxidative (124) decomposition, as well as in the study of many advanced material applications (125). However, they are highly parametrized for specific use cases and are not as accurate as QM methods.

Moreover, when considering the study of larger biomolecular systems, such as proteins and nucleic acids, at larger time scales, where chemical reactions are not the main focus, a classical approximation of the quantum properties of molecules is often sufficient. This is where molecular mechanics (MM), a simplified model for describing molecular systems, comes into play. In atomistic MM (from here on out, referred to only as MM), the atoms, considered the smallest unit, are treated as spheres with a certain mass and point charge bonded through springs. MM is useful in describing the spatial rearrangement of atoms, as in the study of protein dynamics and protein-ligand interactions. It is especially useful in molecular dynamics (MD) simulations, which shall be described soon. However, to solve Newton's equations of motion that encompass MD, we first have to calculate the potential energy of all atoms of the system. To do that, we need a force field.

2.1. Force Fields

The term force field refers to the functional form (i.e., equation) and set of parameters used to calculate the potential energy function of a system. For the study of biomolecules, some of the most widely used all-atom force fields are the AMBER force fields, initially developed by Peter Kollman and his co-workers in 1994 (denoted as ff94 (126)) and since progressively updated until the current iteration, ff19SB (127). The previous iterations, ff99SB (128) and ff14SB (129) are, to this day, some of the most used AMBER force fields developed for the simulation of proteins. Since then, several others have been developed

to simulate nucleic acids, such as DNA (e.g., OL15 (130)) and RNA (e.g., OL3 (131)), carbohydrates, and lipids. The General AMBER force field (GAFF) was also developed with a focus on organic compounds, which is the case for most protein ligands (132,133). However, while all these force fields have their intricacies, they all follow the same general equation.

The general force field equation gives us the total energy ($U(r)$) of the system and comprises two types of additive terms: the bonded and non-bonded interaction terms (**Equation 1**) (134–136).

$$U(r) = U_{bonded}(r) + U_{non-bonded}(r)$$

Equation 1. General force field equation representing the total energy of a system treated with MM.

2.1.1. Bonded Interactions

The potential energy function for the bonded interactions is normally split into three terms: bond stretch, bond angle bend, and dihedral torsion. Nevertheless, some force fields include extra terms like the improper dihedral torsion which is a special type of dihedral interaction (**Equation 2**). Others, like the AMBER force fields, handle these terms using the same formulation as the dihedral torsion terms.

$$U_{bonded}(r) = \sum_{bonds} U_{bond}(l) + \sum_{angles} U_{angle}(\theta) + \sum_{dihedrals} U_{dihedral}(\phi) + \left(\sum_{impropers} U_{improper}(\psi) \right)$$

Equation 2. Potential energy function for the bonded interactions term of a general biomolecular force field, consisting of bond stretch, bond angle bend, and dihedral torsion. Some force fields also include a fourth term for improper dihedral torsion.

In biomolecular force fields, typically, both the bond stretch and angle bend terms use a harmonic oscillator approximation obeying Hooke's law, which states that the force needed to compress or extend a spring is proportional to its change in length. In this context, the "spring" represents the bond between two atoms, meaning that every bond has a specific "stiffness" denoted by the force constant k_l . Consequently, the potential

energy for the bond length displacement (l) relative to its equilibrium position (l_0) is given by **Equation 3**.

$$U_{bond}(l) = \frac{1}{2}k_l(l - l_0)^2$$

Equation 3. Energy function for the bond term. k_l : force constant, l : bond length, l_0 : equilibrium bond length.

A similar approach is used to calculate the potential energy of bond angle bending (**Equation 4**). This term also follows a harmonic potential and requires two parameters: the force constant k_θ and the equilibrium bond angle value θ_0 . The potential energy increases relative to how far the angle displacement θ is to its equilibrium value.

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

Equation 4. Energy function for the angle term. k_θ : force constant, θ : bond angle, θ_0 : equilibrium bond angle.

So far, both the potential energy of the bond length stretching and the bond angle bending were calculated similarly. However, the computation of the potential energy of dihedral angles is different. Consider the imaginary atoms A, B, C, and D connected in sequence. The potential energy of the torsion angle term is given by a truncated Fourier series correspondent to the movement of the plane ABC relative to plane BCD (**Equation 5**).

$$U_{dihedral}(\phi) = \sum_{n=1}^N \frac{1}{2}V_n(1 + \cos(n\phi - \gamma_n))$$

Equation 5. Energy function for the dihedral term, which is described by an N-term truncated Fourier series. V_n : dihedral barrier height, n : periodicity, ϕ : dihedral angle, γ_n : dihedral phase.

The improper torsion angle term, also defined by four atoms, is used in some force fields to maintain tetrahedral or planar geometry. Consider the same atoms as before, but this time B is the central atom to which all the others are connected. The improper dihedral angle is the angle between the planes ABC and ACD. This term also follows a harmonic potential in which ψ corresponds to the angular deviation from an equilibrium angle ψ_0 (**Equation 6**). In practice, this term penalizes any atom that bends out-of-plane (e.g., in aromatic rings) or tries to invert its chirality.

$$U_{improper}(\psi) = \frac{1}{2}k_{\psi}(\psi - \psi_0)^2$$

Equation 6. Energy function for the improper dihedral term. k_{ψ} : force constant, ψ : improper torsional angle, ψ_0 : equilibrium improper torsional angle.

Figure 12 summarizes all the bonded interactions accounted for in the potential energy expression of a typical biomolecular force field.

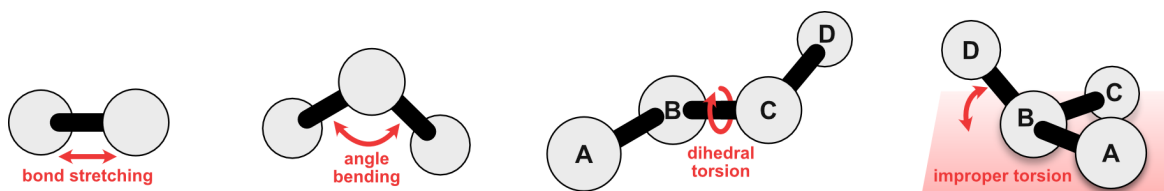


Figure 12. Illustration of all the bonded interactions considered in MM force fields used in the study of biomolecular systems.

2.1.2. Non-Bonded Interactions

Now that we have looked at the bonded interactions accounted for the potential energy expression of a system treated using MM, we must look at the non-bonded interactions. These interactions are usually calculated only for atoms separated by three or more bonds or that belong to different molecules.

Non-bonded interactions can be split into two terms: the van der Waals and the electrostatic interactions (**Equation 7**). The van der Waals component of this term is computed by a 12-6 Lennard-Jones potential, containing both an attractive and repulsive term. On the other hand, the electrostatic interactions are described by Coulomb's law, in which atom-centred point charges represent the charge distribution.

$$U_{non-bonded}(r) = \underbrace{\sum_{i < j} 4 \varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]}_{\text{van der Waals}} + \underbrace{\sum_{i < j} \frac{q_i q_j}{4\pi \varepsilon_0 r_{ij}}}_{\text{electrostatic}}$$

Equation 7. Potential energy function for the non-bonded interactions term of the general force field. i, j : atom pair, ε_{ij} : potential well depth, σ_{ij} : interatomic distance at the energy minimum, r_{ij} : interatomic distance of atoms i and j , q_i and q_j : point charge of the atoms i and j , respectively, ε_0 : electric constant.

Finally, having the complete potential energy expression, it is possible to calculate the total energy of a biomolecular system in a fraction of a second. To do so, we need the coordinates of all atoms of the system, information about how those atoms are connected (i.e., their topology), and the empirical parameters provided by the force field of choice for all the energy terms on the expression (i.e., the force constants, equilibrium distances/angles, atomic charges, and van der Waals parameters).

2.2. Energy Minimization

Before running a molecular dynamics simulation, we must perform an energy minimization of our structure. As previously mentioned, during the preparation of our model, the LeaP software adds any missing heavy atom, protonates the system, adds solvent molecules, and finally adds ions. However, during any of these steps, some steric clashes may occur in the system. This is why minimizing the system is so important, allowing the structure to relax and eliminate any possible clash. Failing to do so may result in the simulation crashing. Moreover, energy minimization also brings the system closer to the nearest local minimum of the potential energy surface, as it is very far from equilibrium at the start.

In a typical MM energy minimization, two algorithms are commonly used: the steepest descent and the conjugate gradient. Both have their pros and cons. While the steepest descent algorithm is considerably faster than the conjugate gradient, it requires more iterations to achieve the desired energy threshold. On the other hand, the conjugate gradient is slower but reaches an energy minimum in a lower number of iterations. Therefore, we start every minimization procedure with the steepest descent algorithm to rapidly approximate our minimum well, then switch to the conjugate gradient algorithm in the final part of the minimization. With this strategy, we can achieve a good balance between accuracy and performance.

In order to progressively adjust the position of all atoms of the system without causing further clashes, the energy minimization procedure is divided into four consecutive stages. Firstly, all the water molecules of the solvent are optimized while the rest of the system is restraint. Then, the position of all hydrogen atoms of the system is minimized, followed by a minimization of all residues sidechains. Lastly, the whole system is minimized until the desired threshold.

2.3. Molecular Dynamics Simulations

A molecular dynamics (MD) simulation is, in simple terms, a common computational approach used to predict the evolution of a system throughout time. In our case, the system consists of a set of atoms that compose a larger biomolecular system, such as a protein or nucleic acid, plus the surrounding solvent molecules and ions. This type of simulation provides us with useful information about its conformational space, thermodynamics, and structural flexibility.

In classical MD simulations, most of the computational effort is spent calculating the gradient ∇ with respect to each atom's position in the x, y, and z directions, which gives us the force acting upon those atoms, as shown in **Equation 8**.

$$\mathbf{F}_i = -\nabla_{\mathbf{r}_i} U(\mathbf{r}_1, \dots, \mathbf{r}_N) = -\left(\frac{\partial U}{\partial x_i}, \frac{\partial U}{\partial y_i}, \frac{\partial U}{\partial z_i}\right)$$

Equation 8. Mathematical relation between the force $\mathbf{F}_i$ acting upon particle i , and its potential energy U . $U(\mathbf{r}_1, \dots, \mathbf{r}_N)$ represents the potential energy of N interacting atoms as a function of their positions $\mathbf{r}_i = (x_i, y_i, z_i)$.

In MD simulations, the time evolution of a set of interacting particles is followed via the solution of Newton's equations of motion, described in **Equation 9**, where $\mathbf{r}_i(t) = (x_i(t), y_i(t), z_i(t))$ is the position vector of particle i with mass m_i and $\mathbf{F}_i$ is the force acting on that particle at time t .

$$\mathbf{F}_i = m_i \frac{\partial^2 \mathbf{r}_i(t)}{\partial t^2}$$

Equation 9. Second order derivative of Newton's second law. $\mathbf{F}_i$: force acting on particle i , m_i : mass of the particle i , $\mathbf{r}_i$: position vector of particle i , t : time.

However, it is impossible to analytically solve Newton's equations of motion for a system with more than two atoms, a problem known in physics as the n -body problem. Therefore, we must use a numerical method to integrate Newton's equations of motion. The most popular integration methods in MD simulations are the Verlet algorithm (137) and its variants, the Leapfrog integrator, and the velocity Verlet algorithm (138).

The aim of numerically integrating Newton's equations of motion is to find an expression that defines positions $\mathbf{r}_i(t + \Delta t)$ at time $t + \Delta t$ in terms of already known positions at time t . This can be achieved by the Verlet algorithm (**Equation 10**).

$$\mathbf{r}_i(t + \Delta t) \cong 2\mathbf{r}_i(t) - \mathbf{r}_i(t - \Delta t) + \frac{\mathbf{F}_i(t)}{m_i} \Delta t^2$$

Equation 10. Basic equation of the Verlet algorithm derived from the Taylor expansions for the positions $\mathbf{r}_i(t)$. This equation is accurate to an order given by the fourth power of the time step Δt .

Velocities can then be calculated from the positions or be propagated explicitly, as in Leapfrog and velocity Verlet algorithms. Updating the atomic positions and velocities in a stepwise procedure gives us their trajectory. While the exact trajectories correspond to the limit of an infinitesimally small integration step, in practice, we use larger time steps. This aspect will be discussed in the following section.

2.3.1. Integration Time Step

To numerically integrate Newton's equations of motion, using any of the algorithms mentioned above, we need to define a time step (Δt) for the simulation, which is determined by the fastest frequency motion in the system. Since during an integration time step, the velocity and acceleration felt by all the system's atoms should be approximately constant, using a larger time step would lead to aberrant movements, likely crashing the simulation. In the case of the simulation of a biomolecular system, the vibration of C-H bonds is usually the fastest. Therefore, to provide a good description of any atom's movement, a good rule of thumb is to use a time step at least ten times smaller than the fastest bond vibration, which is 11 fs for C-H bonds. Thus, the recommended time step for this case is 1 fs.

However, using 1 fs time step is computationally demanding, making the simulation of bigger systems or larger time frames unsurmountable. As hydrogen atoms' dynamics are usually not essential for a simulation, bonds involving hydrogens (e.g., C-H, O-H, N-H,...) can be constrained. Replacing bond vibrations with holonomic (i.e., unchanging) constraints allow doubling the simulation step to 2 fs since the following fastest motions (i.e., bond vibrations involving only heavy atoms and angles involving hydrogen atoms) have a period of around 20 fs. If bonds between all atoms and angles involving hydrogen atoms are constrained, the next fastest bond vibration will have a period of 45 fs, increasing the maximum time step to around 4 fs.

Constraint algorithms like SHAKE (139) and LINCS (140) are commonly used to achieve this. Moreover, the SETTLE algorithm (141), an analytical solution of SHAKE, is also typically used in conjunction with SHAKE or LINCS to constrain rigid water molecules

(e.g., TIP3P or SPC/E water models), which often make up more than 80% of the simulation of a typical biomolecular system. SETTLE and SHAKE are available in AMBER and GROMACS, while LINCS, considered to be faster and more stable than SHAKE, is only available in GROMACS. What these algorithms do in practice is constraint bond distances removing the bond stretching freedom of those bonds.

Another common way to increase the time step to 4 fs, speeding up the simulation, is by performing hydrogen mass repartitioning (HMR) (142). HMR redistributes some of the mass of heavy atoms to all their bonded hydrogens, maintaining the same overall mass. This allows for longer time steps by slowing the highest-frequency motions of the macromolecule under study.

2.3.2. *Ensembles*

Statistical mechanics describe how molecular events and macroscopic properties are related. In statistical mechanics, an ensemble, more specifically a thermodynamic ensemble, is used to derive the properties of thermodynamic systems from the laws of classical MM.

Some thermodynamic properties like temperature, pressure, volume, and particle number, among others, are examples of state functions. State functions quantitatively describe an equilibrium state of a thermodynamic system, regardless of how the system has arrived at that state. A macroscopic system can be described by an ensemble that depends only on a few observable parameters. In the study of biomolecular systems, the most important ensembles are the canonical ensemble (NVT), the isothermal-isobaric ensemble (NPT), the microcanonical ensemble (NVE), and the grand canonical ensemble (μ VT).

Simulations of biomolecular systems are typically conducted in the Isothermal-Isobaric (NPT) ensemble, as it is the one that better resembles physiological conditions, where pressure and temperature are usually constant. Moreover, the notion of Gibbs free energy is only meaningful when changes occur at a constant temperature and pressure. However, a short simulation in the NVT ensemble is necessary to equilibrate the system to the desired temperature. Therefore, our studies used both ensembles: NVT in the equilibration stage and NPT in the production stage.

2.3.3. *Temperature and Pressure Control*

The simulation of a biomolecular system is usually carried out using the NPT ensemble, where the following state functions — number of particles (N), pressure (P), and temperature (T) — must be constant. Since the number of particles in the system is immutable, this condition is assured throughout the simulation. On the other hand, to achieve temperature and pressure control, different algorithms must be used. There are several algorithms, but the ones used in this work are the Langevin thermostat (143–145) for temperature control and the Berendsen barostat (146) for pressure control.

2.3.4. *Periodic Boundary Conditions*

Why do we need to use periodic boundary conditions? A molecular system studied with MD simulations cannot be dissociated by the environment surrounding it, which in the case of biomolecular systems usually means an aqueous solution.

Therefore, there are two ways that we can use to account for the presence of the solvent. A simpler approach is to use an implicit solvation model to simulate the solvent's and counterions' presence. This approach adjusts the dielectric constant to mimic the effect of the solvent on the electrostatic interactions and adds empirical coefficients to simulate the solvent's viscosity. Another alternative is to include water molecules and counterions in the simulation explicitly. However, while this approach is better, it also introduces a new problem—the description of intermolecular interactions at the solvent box's boundary. If a single unit cell were to be used, all the water molecules would “evaporate” into oblivion. It would be like simulating a droplet of water in a vacuum. Therefore, to simulate a realistic environment, preserving thermodynamic properties like temperature, pressure, and density, periodic boundary conditions (PBC) are used. This approach provides a way to study a system on a bulk solvent without significantly impacting the computational cost. PBC is implemented by surrounding the unit cell (also known as the periodic box) by translating copies of itself in all directions, thereby approximating an infinitely large system. The motion equations are only solved on the central unit cell and are replicated in all the others. This means that when one molecule diffuses across the boundary of the simulation box, it will reappear on the opposite side of the box with the same velocity, thereby keeping the number of particles inside the simulation box constant.

However, certain considerations must be taken into account to use the PBC approach successfully. For example, interactions of a molecule with its image on a neighbour replica

cell are not desirable. To overcome this, the minimum distance between the biomolecule of interest and any face of the water box should be larger than the cutoff applied for the non-bonded interactions (described in detail in the next section). For example, if the non-bonded interactions cutoff applied is 10 Å, a reasonable minimum distance would be 12 Å.

2.3.5. *Non-Bonded Interactions Cutoff*

Non-bonded interactions are calculated as a sum of electrostatic and non-electrostatic interactions. However, due to the large number of non-bonded interactions to compute, this part of the calculation can quickly become a bottleneck. As a result, the calculation of the non-bonded interactions is often the most time-consuming part of an MD simulation.

For this reason, some schemes have been devised to make this process more efficient. One of these schemes is the application of a radial cutoff that limits the accounted interactions of each atom. To avoid problems caused by an abrupt discontinuation in the potential, a switching function is typically applied to reduce the potential energy contribution and its resultant forces smoothly to zero. As the effect of the non-bonded interactions gets smaller with distance, this approximation is a good balance between accuracy and computational cost. However, a reasonable cutoff value must be chosen to avoid leading to aberrant results.

2.3.6. *Analysis of the Molecular Dynamics Simulation*

MD simulations are powerful tools, providing lots of helpful information for the study of the catalytic mechanism of an enzyme. However, when performing an MD simulation, it is essential to verify that some quality-indicating measures like density, volume, pressure, temperature, kinetic energy, and potential energy behave normally throughout the simulation. Here “quality” refers to the overall thermodynamic stability of the simulation and is not necessarily an indication of its accuracy in reproducing the physical phenomena of interest.

Moreover, measures like root-mean-square deviation (RMSD), radius of gyration (R_g), and root-mean-square fluctuation (RMSF) are useful to indicate the convergence and fluctuation of a system. The RMSD of a biomolecular system over time can be used to assess how much its conformation changes during a simulation, providing an indication of conformational stability. For a converging simulation, we expect the RMSD to increase

and then plateau, even though the plateau of RMSD alone does not necessarily mean that a system has converged. One should look to all quantities of interest, i.e., any measures relevant to the behaviour and dynamics of our system, and determine if those have converged. A helpful metric can be the R_g , which indicates how the mass of a group of atoms is distributed around their centre of mass. In a converged trajectory, the R_g should also reach a plateau. Lastly, the RMSF measures the fluctuation of atoms around their average position. Any large fluctuations should be compared to the crystallographic B-factors.

Furthermore, analysis like the identification and time-distribution of hydrogen bonds between a protein and a ligand or a substrate and the residues at an enzyme's active site should also be considered. Another handy measure is the radial distribution function (RDF) of water, which gives us the number of water molecules as a function of distance. These types of analysis allow us to obtain essential information, primarily regarding the active site interactions between the ligand and the enzyme.

2.4. Molecular Docking

Molecular docking is a computational approach used to predict the most favourable ligand-target configuration(s) and estimate the strength of the interaction(s) between the ligand and the target. In practical terms, docking is used to study how two molecules interact. They can be a small ligand, a protein, DNA, or RNA, among others. The most commonly studied complexes are protein-ligand and protein-protein.

There are innumerable small-ligand docking software, like AutoDock (147), AutoDock Vina (148,149), LeDock (150), GOLD (151), MOE (152), and Glide (153–155). However, this work focuses its attention on protein-nucleic acids docking, and one of the most used software for studying these complexes is HADDOCK (156–158).

Therefore, in the following subsections, I will briefly describe the different types of docking, the algorithms used, and the scoring functions used to rank the conformations obtained. Then, I will delve more into HADDOCK, what type of docking software it is, and how it works. Moreover, I will talk about binding free energy calculations using MM-GBSA/MM-PBSA as a way to refine the solutions obtained with molecular docking (159).

2.4.1. Search algorithms

To dock a ligand into its binding site, several degrees of freedom need to be computed. There are three types of docking algorithms: rigid, semi-flexible, and flexible algorithms (135). These are defined based on the ligand's and target's flexibility.

The rigid docking algorithms were the first to be implemented in molecular docking and are typically used in the first stage of docking by many software (including HADDOCK). In rigid docking, the ligand and the target are treated as rigid bodies. For this reason, only six degrees of freedom are considered, three translational and three rotational, that define the position and orientation of the ligand relative to its target.

Unlike the rigid docking approach, semi-flexible docking algorithms consider the ligand's flexibility. However, since it is computationally expensive to explore all the possible conformations of the ligand, especially for increasing ligand sizes, several approaches were developed to address the problem of ligand sampling. They are classified as systematic, stochastic, and deterministic. The systematic search algorithms explore the ligand's degrees of freedom incrementally. On the other hand, stochastic search algorithms perform random changes in the ligand's degrees of freedom based on an evolutionary strategy, like the genetic algorithm or the tabu search method. This type of search algorithm is the most common, used by software like AutoDock and GOLD. Finally, the deterministic search algorithms use energy minimization methods and molecular dynamics simulations to obtain a ligand conformation and orientation with a low energy value. This type of algorithm, however, presents a higher computational cost and could lead to the ligand being trapped in an undesirable local energy minimum.

Lastly, flexible docking algorithms are algorithms where the protein conformation space is also explored. There are several ways to account for target flexibility. Some ways are soft docking, side-chain flexibility (typically using side-chain rotamer libraries), molecular relaxation (using Monte Carlo or MD simulations), or ensemble docking.

2.4.2. Scoring Functions

A scoring function is a mathematical expression used to evaluate the binding affinity between a ligand and its target or between two biomolecules (e.g., protein-protein and protein-DNA).

There are three types of scoring functions: force-field-based, empirical, and knowledge-based (135). The force field-based scoring functions are a sum of intra- and inter-molecular contributions, just those described in **section 2.1**. Molecules are represented using force field-specific atom and bond types with atom-centred partial charges. The bonded interactions derive from the bond stretching, bond angle, dihedral angle, and improper dihedral angle energy terms. At the same time, the non-bonded interactions correspond to the electrostatic and the van der Waals energy terms, described by Coulomb's law and a 12-6 Lennard-Jones potential, respectively. These types of scoring functions are usually more computationally demanding than the rest. A notorious example is the AMBER scoring function used in DOCK software (160,161).

On the other hand, the empirical scoring functions estimate the binding free energy as a sum of uncorrelated energy terms. Its coefficients are obtained by regression analysis of a set of ligands with known experimental binding free energy. However, their accuracy depends on how well the ligand and receptor are represented in the training dataset used to fit those coefficients. Two very known examples of empirical scoring functions are the ChemScore (162,163) and the Piecewise Linear Potential (PLP) (164–167) used in GOLD.

Lastly, the knowledge-based scoring functions are composed of multiple weighted molecular features related to the ligand-receptor binding modes. One example is the Astex Statistical Potential (ASP) scoring function used in GOLD (168).

2.4.3. HADDOCK

HADDOCK (High Ambiguity Driven protein-protein DOCKing) is a flexible docking approach for the modelling of biomolecular complexes (156–158). While it was initially developed for protein-protein docking, HADDOCK can also model protein-nucleic acids (169–178) and protein-ligand (179–184) complexes. It distinguishes itself from other *ab initio* docking methods by being information-driven. However, if no information about the binding site is available, it randomly defines ambiguous interaction restraints (AIRs) from solvent-accessible residues to drive the docking process. This approach, where a ligand is docked without prior knowledge of the target binding site, is called blind docking.

HADDOCK workflow consists of a series of three docking stages. The first one, it0, consists of a rigid body docking where the ligand orientation is randomized relative to the protein to optimize their interaction. The second is a semi-flexible refinement (it1), where a molecular dynamics-based refinement introduces ligand and protein flexibility to optimize

the residues at the interface. Lastly, there is a refinement of the complex obtained in explicit solvent (water).

From this, scoring is performed according to the weighted sum (HADDOCK score) shown in the following equation (185,186):

$$HS = 1.0 E_{vdW} + 0.2 E_{elec} + 1.0 E_{desolv} + 0.1 E_{AIR}$$

Equation 11. HADDOCK scoring function. E_{vdW} : van der Waals intermolecular energy, E_{elec} : electrostatic intermolecular energy, E_{desolv} : desolvation energy, E_{AIR} : distance restraints energy.

The HADDOCK score, contrarily to the scoring functions previously mentioned, is a combination of energy terms calculated with the MM force field OPLS (i.e., vdW and electrostatic interactions), empirical terms (i.e., desolvation and buried surface area), and restraint violation energies (E_{AIR}).

However, this general equation is not valid for all situations. For example, in the case of protein-nucleic acids docking, HADDOCK recommends setting the E_{desolv} weight to 0.

Finally, the structures are ranked according to their score, and the one with the smallest value is ranked first.

2.5. Binding Free Energy Calculations

Molecular docking is a great way to predict possible protein-ligand interactions, be it another protein, DNA, RNA, or a small ligand. Therefore, it is an instrumental method to predict binding modes and distinguish between binders and non-binders. However, it cannot properly estimate binding free energies. This is where MM-PBSA and MM-GBSA (molecular mechanics energies combined with the Poisson-Boltzmann or Generalized Born and Surface Area continuum solvation) methods come into play (159). These are two popular end-point methodologies used to estimate the binding free energy of a ligand to a receptor. Although more computationally expensive than docking, they are much less expensive than the more accurate alchemical perturbation methods. Therefore, they present a good balance between accuracy and computational cost, especially for relatively large ligands, such as DNA or RNA molecules.

MM-PB(GB)SA is a post-processing method in which snapshots from an MD trajectory are used to calculate the free energy change between two states, the bound and free state of a receptor and ligand. Two protocols can then be used to obtain those snapshots: a 1-

trajectory or a 3-trajectory protocol. Whereas the 3-trajectory protocol requires the simulation of the complex, the free ligand, and the unbound receptor independently, the 1-trajectory protocol only requires the simulation of the complex. Therefore, every snapshot for the complex, ligand, and receptor is extracted from the same trajectory file. On the other hand, this approach considers that there is no significant change in the structure of the receptor and the ligand upon its binding, which may be an essential aspect for its affinity.

Since the binding free energy of the solvated complex is very hard to calculate directly, it is estimated from the free energies of the complex, free protein, and ligand, as shown in the following equation.

$$\Delta G_{bind} = \langle G_{PL} - G_P - G_L \rangle_{PL}$$

Equation 12. The binding free energy (ΔG_{bind}) is estimated from the free energies of the complex (PL), protein (P), and ligand (L). However, when using the 1-trajectory protocol, only the complex is simulated. The ensemble average of the free receptor and ligand is created simply by removing the respective atoms.

In MM-PB(GB)SA, the free energy (G) of a state (i.e., P, L, or PL) is estimated as follows:

$$G = \underbrace{E_{bnd} + E_{el} + E_{vdW}}_{\text{gas phase}} + \underbrace{G_{pol} + G_{np}}_{\text{solvation}} - TS$$

Equation 13. The free energy is calculated by combining the so-called gas phase energy contributions (i.e., E_{bnd} : bonded interactions energy, E_{el} : electrostatic interactions energy, and E_{vdW} : van der Waals interactions energy) with the solvation free energy components (G_{pol} : polar contribution to the solvation free energy, and G_{np} : non-polar contribution to the solvation free energy). In addition, entropy contributions may be added as a further refinement (TS).

The first three terms are standard MM energy terms from bonded (i.e., bond, angle, dihedral, and improper dihedral), electrostatic, and van der Waals interactions. G_{pol} and G_{np} are the polar and non-polar contributions to the solvation free energies, respectively. G_{pol} is calculated using either the PB method (MM-PBSA approach) or the GB method (MM-GBSA approach). The GB method is a faster and more efficient approximation of the PB method, as it provides an analytical expression for the calculation of the polar contribution to the solvation free energy. Therefore, it was the preferred method used in our works, especially using the *GB-neck2* model (187) (*igb* = 8 in AMBER), which like the model from which it was based (i.e., *GBn* (188)), yields results in agreement with the more computationally demanding PB method, as well as, explicit solvent simulations (189).

However, contrary to the *GBn* model, *GB-neck2* was refined for systems involving nucleic acids, being the currently recommended choice (187). G_{np} , on the other hand, is estimated from a linear relation to the solvent-accessible surface area (SASA), typically estimated with the LCPO (Linear Combinations of Pairwise Overlaps) method (190).

3. Quantum Mechanics

For centuries the classical mechanics theory developed by Sir Isaac Newton was instrumental in describing the motion of everyday objects and large bodies (e.g., planets). However, by the end of the nineteenth century, scientists started to make observations for which classical mechanics had no explanation. They were therefore compelled to revise their theory to describe the observed phenomena at very small scales – then, quantum mechanics (QM) was born.

While classical MD, which uses molecular mechanics, is a great and powerful methodology to study the time evolution of biomolecular systems at a low computational cost, it cannot describe chemical processes that involve electrons, such as bond breakage and formation. In order to study these phenomena, we need quantum mechanics, namely *ab initio* methods.

Although QM methods are much more accurate, they present a disadvantage, they cannot be applied directly to systems with more than a few hundred atoms, or it becomes computationally too expensive. In the case of enzymes, the part where the reaction occurs, the active site, is the only part we need to treat with QM methods. Therefore, we are left with two choices. We either keep only the active site residues essential to the catalysis and discard the rest, the so-called QM cluster model. Alternatively, keep the entire enzyme intact but only treat those active site residues with QM, while the rest is treated with MM. This so-called hybrid method will be discussed in more detail in **section 4**.

Nevertheless, what is QM, and how can it be used to study chemical reactions? The use of QM involves the application of the wavefunction to the electronic configuration of the system. However, due to the many-electron problem, the wavefunction equation cannot be solved analytically. For this reason, several new methods have been developed throughout the years to try to overcome this limitation. They all try to provide a good balance between accuracy and computational cost.

The first wavefunction-based method developed for the study of many-electron systems was the Hartree-Fock (HF). However, HF presents a problem. It does not accurately account for electron correlation, a critical flaw responsible for HF's inability to describe London dispersion forces.

For this reason, several so-called post-HF methods were proposed to overcome this limitation, particularly the Møller-Plesset (MP) perturbation theory and the Coupled cluster

(CC). Unfortunately, these methods usually give more accurate results at the expense of being more computationally demanding, which makes them only really useful for small systems.

In the 70s, a new method was developed that was not based on the wavefunction equation but on the electron density of the system. Its name is Density Functional Theory (DFT). One of DFT's greatest strengths is that it provides a great balance between speed and accuracy. Currently, several exchange-correlation DFT functionals try to apply the best approximation to the exchange-correlation energy since its exact form is unknown.

These QM methodologies will be discussed in more detail in the following sections of this chapter.

3.1. Energy Quantization

In physics, energy quantization is the phenomenon that energy exists in discrete values. This means that energy can only be transferred or emitted in certain amounts, called quanta. The discovery of energy quantization led to the development of quantum mechanics.

Three experiments performed by the end of the nineteenth century led scientists to believe that energy can only be transferred in discrete amounts. The first of these experiments concerns blackbody radiation.

An ideal body capable of absorbing and emitting light at all frequencies is called a blackbody. Consequently, the radiation emitted by a blackbody is called blackbody radiation. The work from Lord Rayleigh and J. H. Jeans culminated in the Rayleigh-Jeans law, an equation that tried to explain the blackbody radiation and approximates the experimental data to longer wavelengths. However, as frequencies increased (and wavelengths decreased), the Rayleigh-Jeans law failed, a phenomenon named ultraviolet catastrophe.

Only in 1900, the German physicist Max Planck was able to explain blackbody radiation (191). Planck's hypothesis states that energy is emitted and absorbed in discrete packets, or quanta. This hypothesis was developed to explain the ultraviolet catastrophe, which is the problem that classical physics has in explaining why hot objects do not emit an infinite amount of high-frequency radiation. According to classical physics, the hotter an object is, the more high-frequency radiation it should emit. However, experiments showed that hot objects do not emit infinite high-frequency radiation. Planck's hypothesis solved this

problem by proposing that energy is emitted and absorbed in discrete packets, which means that there is a limit to the amount of high-frequency radiation that a hot object can emit. In other words, he found that energy is quantized and that this relation depends on a constant, later honoured in his name as Planck's constant.

$$E_{\text{photon}} = h\nu$$

Equation 14. Relation between the energy of a photon (E_{photon}) and its frequency (ν). h : Planck's constant.

3.2. Wave-Particle Duality

Later Einstein applied Planck's hypothesis to explain the photoelectric effect by proposing that light, considered a wave, is composed of particle-like entities, which he named photons (192). In this phenomenon, electrons are ejected from a metal surface when exposed to light. Furthermore, Einstein found that the energy of the ejected electrons was always a multiple of a certain amount.

Then, in 1924, Louis de Broglie proposed that particles, such as electrons, can exhibit both wave-like and particle-like behaviour, analogous to the behaviour of light explained by Einstein. According to his equation, the motion of an electron of mass m and velocity v has a wavelength λ .

$$\lambda = \frac{h}{p} = \frac{h}{mv}$$

Equation 15. De Broglie's equation. λ : wavelength; h : Planck's constant; p : momentum of the particle; m : mass of the particle; v : particle's velocity.

His theory was later confirmed experimentally with the observation of electron diffraction.

The wave-particle duality is one of the most important principles in quantum mechanics, providing an explanation for Bohr's model of the atom. Moreover, this principle has led to the development of many important concepts, such as Heisenberg's Uncertainty Principle (193).

3.3. Heisenberg Uncertainty Principle

The Heisenberg Uncertainty Principle (193) is one of physics' most famous and widely-known principles. Proposed by Heisenberg in 1927, it states that it is impossible to know both the position and momentum of a particle at the same time. Therefore, the more accurately you know one, the less accurately you can know the other. This principle is fundamental to the understanding of quantum mechanics.

$$\Delta x \Delta p \geq \frac{\hbar}{2}$$

Equation 16. Heisenberg's Uncertainty Principle. Δx : uncertainty in position; Δp : uncertainty in momentum; $\hbar$: reduced Planck constant, which is equal to the Planck constant h divided by 2π .

All these works showed that classical mechanics could not be used to describe microscopic phenomena. Nevertheless, how can we use quantum mechanics to describe such a system? Here enters the Schrödinger equation and the concept of the wavefunction.

3.4. The Schrödinger Equation

The wavefunction is a mathematical function that describes the matter and energy's wave-like behaviour. In quantum mechanics, the wavefunction is used to describe the behaviour of particles at the atomic and subatomic levels. For example, the wavefunction can be used to determine properties such as position, momentum, and energy.

By the end of 1925, after giving a presentation about de Broglie's work, Peter Debye pointed out to Schrödinger that if such a wave exists, there ought to be a wave equation that describes it. According to many, this pivotal moment led Schrödinger to discover the equation that is now forever linked to his name.

The Schrödinger equation is a linear partial differential equation that describes the wavefunction of a quantum mechanical system (194). However, it is impossible to solve the Schrödinger equation analytically for a system larger than the hydrogen atom.

If we consider a three-dimensional system consisting of one particle, its evolution is given by the following equation:

$$-\frac{\hbar^2}{2m}\nabla^2\Psi(r,t) + V(r,t)\Psi(r,t) = i\hbar\frac{\partial}{\partial t}\Psi(r,t)$$

Equation 17. Three-dimensional time-dependent Schrödinger equation.

Where ∇^2 , the Laplacian operator, is defined as:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

Equation 18. Laplacian operator.

However, a more simple, one-dimensional system can be described by the following equation:

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\Psi(x,t) + V(x,t)\Psi(x,t) = i\hbar\frac{\partial}{\partial t}\Psi(x,t)$$

Equation 19. One-dimensional, time-dependent Schrödinger equation.

Nevertheless, in most applications of the Schrödinger equation, it is common to use its time-independent form, which the following equation can describe. Only problems such as studying excitation processes require using the time-dependent equation.

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi(x) + V(x)\psi(x) = E\psi(x)$$

Equation 20. One-dimensional, time-independent Schrödinger equation.

The Schrödinger equation can be further simplified into an eigenvalue equation. In such case, the wavefunction is an eigenfunction of the Hamiltonian operator ($\hat{H}$) with eigenvalue E as seen in the following equation:

$$\hat{H}\Psi(r) = E\Psi(r)$$

Equation 21. Reduced Schrödinger equation.

In simpler words, it means that the Hamiltonian ($\hat{H}$) multiplied by the wavefunction ($\Psi(r)$) will give the same wavefunction multiplied by the energy of the system (E).

In quantum mechanics, a wavefunction describes the quantum state of a set of particles in an isolated system. In other words, it contains all possible information about a system. However, no one knows the wavefunction exactly, not even Schrödinger, who initially

interpreted it as a charge density. As a result, up to this day, there are several interpretations of the wavefunction.

3.4.1. Born's Statistical Interpretation of the Wavefunction

In quantum mechanics, the Born interpretation is an interpretation of the wavefunction Ψ that gives the probability of finding a particle in a given state. It was named after Max Born, who first proposed it in 1926 (195). In the Born interpretation, the wavefunction is seen as a probability amplitude, meaning that the absolute square of the wavefunction $|\Psi|^2$ gives the probability density. The Born interpretation is one of the most commonly used interpretations of the wavefunction, and it is essential for the understanding of quantum mechanics.

3.4.2. The Hamiltonian Operator

A mathematical operator is a symbol that represents a particular operation that can be performed on one or more variables (e.g., a derivative $\frac{\partial}{\partial x}$). In physics, mathematical operators are often used to represent physical quantities such as momentum, energy, and electric field (e.g., kinetic energy $-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2}$).

In quantum mechanics, we have the Hamiltonian operator ($\hat{H}$), a fundamental operator that represents the total energy of a system, including the kinetic and potential energies.

$$\hat{H} = \hat{T} + \hat{V}$$

Equation 22. The Hamiltonian operator $\hat{H}$ is the sum of the kinetic energy operator $\hat{T}$, and the potential energy operator $\hat{V}$.

These two energies can be further divided into:

$$\hat{H} = \hat{T}_e + \hat{T}_N + \hat{V}_{Ne} + \hat{V}_{ee} + \hat{V}_{NN}$$

Equation 23. Decomposition of the Hamiltonian operator ($\hat{H}$) into its different kinetic and potential energy terms. $\hat{T}_e$: electrons' kinetic energy; $\hat{T}_N$: nuclei's kinetic energy; $\hat{V}_{Ne}$: nuclei-electron attraction; $\hat{V}_{ee}$: electron-electron repulsion; $\hat{V}_{NN}$: nuclei-nuclei repulsion.

All these energy terms of the Hamiltonian can be differentiated as follows:

$$\hat{H} = \underbrace{-\frac{1}{2} \sum_{i=1}^N \nabla_i^2}_{\text{KE Elec.}} - \underbrace{\frac{1}{2} \sum_{A=1}^M \frac{1}{M_A} \nabla_A^2}_{\text{KE Nucl.}} - \underbrace{\sum_{i=1}^N \sum_{A=1}^M \frac{Z_A}{r_{iA}}}_{\text{Nucl.-Elec. Attract.}} + \underbrace{\sum_{i=1}^N \sum_{j>1}^N \frac{1}{r_{ij}}}_{\text{Elec.-Elec. Repuls.}} + \underbrace{\sum_{A=1}^M \sum_{B>A}^M \frac{Z_A Z_B}{R_{AB}}}_{\text{Nucl.-Nucl. Repuls.}}$$

Equation 24. Differentiation of the energy terms of the Hamiltonian in atomic units for simplicity. *KE*: kinetic energy; *N*: number of electrons; *M*: number of nuclei; *M_A*: mass of nucleus *A*; *Z*: atomic number; *r_{iA}*: distance between electron *i* and nucleus *A*; *r_{ij}*: distance between electrons *i* and *j*; *R_{AB}*: distance between nuclei *A* and *B*. ∇^2 is the Laplace operator and is equal to $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$.

3.5. Variational Principle

We can only solve Schrödinger's equation analytically for a few cases, such as the "particle in a box", harmonic oscillator, rigid rotor, and the hydrogen atom. For all other cases, the best we can do is solve approximate wave functions for every other possible Hamiltonian that comes up. Approximate solutions are obtained using techniques like the variational method or the perturbation theory, which we shall see in later sections.

Using the variational method, we are able to find the best possible approximation for the ground state energy and wavefunction of a quantum mechanical system where we do not know the exact result. It states that the true ground state of a system (i.e., the most stable state of a system — the one with the lowest energy) is the state that minimizes the expected value of the system's energy. This principle can approximate the behaviour of quantum systems by finding the state that minimizes the expected value of the system's energy. Of notice is that the energy obtained from using the variational principle is always greater than or equal to the energy of the ground state.

$$E_0 \leq E_\phi$$

Equation 25. The Variational Principle states that the energy of any approximate ground state wave function is always greater than or equal to the exact energy of the ground state. *E₀*: exact ground state energy; *E_φ*: approximate ground state energy.

The energy of the trial wavefunction can be calculated as follows:

$$E_{\phi} = \frac{\int \psi_{\phi}^* \hat{H} \psi_{\phi} d\tau}{\int \psi_{\phi}^* \psi_{\phi} d\tau} = \frac{\langle \psi_{\phi} | \hat{H} | \psi_{\phi} \rangle}{\langle \psi_{\phi} | \psi_{\phi} \rangle}$$

Equation 26. The energy of the trial wavefunction can be calculated as the expectation value of the Hamiltonian (numerator) divided by the normalization integral (denominator). E_{ϕ} : approximate ground state energy; ψ_{ϕ}^* : complex conjugate of the approximate ground state wavefunction; ψ_{ϕ} : approximate ground state wavefunction; $\hat{H}$: Hamiltonian acting on ψ_{ϕ} .

3.6. Born-Oppenheimer Approximation

The Born-Oppenheimer approximation is a quantum mechanics technique that simplifies the Schrödinger equation's resolution (196). It is based on the assumption that since $m_{nuclei} \gg m_e$, the dynamics of atomic nuclei and electrons can be separated. Therefore, as the nuclei are much heavier than the electrons, they move slower. There is a minimum of three orders of magnitude mass difference between electrons and nuclei. So given the same amount of kinetic energy, E_{KE} , electrons will move much faster than the nuclei. In other words, from the perspective of the electrons, the nuclei are fixed in place. Whereas from the perspective of the nuclei, the electron positions “update instantaneously”. As such, the Hamiltonian can be split into electronic and nuclear parts that can be solved separately.

$$\hat{H} = \hat{H}_{elec} + \hat{H}_{nucl}$$

Equation 27. Division of the Hamiltonian operator ($\hat{H}$) into electronic ($\hat{H}_{elec}$) and nuclear ($\hat{H}_{nuc}$) Hamiltonians.

The electronic Hamiltonian ($\hat{H}_{elec}$) is given by the sum of the electrons' kinetic energy ($\hat{T}_e$) and the potential energy associated with the nucleus-electron attraction ($\hat{V}_{Ne}$), and the electron-electron repulsion ($\hat{V}_{ee}$).

$$\hat{H}_{elec} = -\frac{1}{2} \sum_{i=1}^N \nabla_i^2 - \sum_{i=1}^N \sum_{A=1}^M \frac{Z_A}{r_{iA}} + \sum_{i=1}^N \sum_{j>i}^N \frac{1}{r_{ij}} = \hat{T}_e + \hat{V}_{Ne} + \hat{V}_{ee}$$

Equation 28. Electronic Hamiltonian ($\hat{H}_{elec}$).

On the other hand, as the nuclei are (temporarily) not allowed to move, their kinetic energy is zero. Additionally, for a given set of nuclei positions, the nuclei-nuclei repulsion

does not change, so it can be described as a constant. So, traditionally, the potential energy term $\hat{V}_{NN}$ can be included in the electronic Hamiltonian $\hat{H}_{elec}$.

$$\hat{H}_{nuc} = -\frac{1}{2} \sum_{A=1}^M \frac{1}{M_A} \nabla_A^2 + \sum_{A=1}^M \sum_{B>A}^M \frac{Z_A Z_B}{R_{AB}} = \hat{T}_N + \hat{V}_{NN}$$

Equation 29. Nuclear Hamiltonian ($\hat{H}_{nuc}$).

Therefore, from now on, we can focus on solving the electronic Hamiltonian to find the ground state, i.e., the Ψ_{elec} that gives the lowest E_{elec} .

3.7. Basis Sets

The electronic wavefunction of a single electron system depends only on the nuclear coordinates (fixed based on the Born-Oppenheimer approximation) and the electron's coordinates. However, to conceptualize the wave function of a multi-electron molecule, we must consider all the orbitals involved. We can then construct a guess of the system's wavefunction by combining atomic orbitals or hybrid orbitals from each atom of the molecule or other molecular orbitals from groups of atoms.

$$\chi_i = \sum_{\mu=1}^K c_{\mu i} \tilde{\chi}_{\mu}$$

Equation 30. Representation of an orbital as the sum of K basis functions. χ_i : orbital i ; K : number of basis functions; $c_{\mu i}$: coefficient; $\tilde{\chi}_{\mu}$: basis function.

Linear Combination of Atomic Orbitals (LCAO) is a quantum chemistry method that combines all the individual atomic orbitals - basis sets - to obtain the molecular orbitals of a system. There is only one requirement here: we must know the wave function of all the atomic orbitals of the system. Since Schrödinger's equation can only be solved exactly for hydrogen and hydrogen-like systems, it is imperative to determine the equation for the remaining atoms.

Different families of functions have been developed to overcome this problem in order to reproduce the properties required for an atomic orbital. In addition, different basis sets provide varying degrees of accuracy in the description of orbitals, resulting in different computational costs.

In theory, the system would have the most accurate wave function if there were an infinite number of basis sets. Nevertheless, it is not possible to compute an infinite number of functions. Since only a limited number of functions are used, the basis set is truncated, resulting in additional error. Hence, our objective is to minimize the number of basis functions while maintaining a reasonable level of accuracy.

The Slater Type Orbitals (STO) and the Gaussian Type Orbitals (GTO) were two of the first basis sets developed for studying molecular systems.

3.7.1. Slater and Gaussian Type Orbitals

There are two main types of atomic orbitals, the Slater Type Orbitals (STO) and the Gaussian Type Orbitals (GTO).

For an atom with atomic number Z , the STO is described as the function of the effective nuclear charge Z_e and the quantum numbers associated with the orbital: n , l , and m_l .

$$\chi_{STO} = NY_{l,m}(\theta, \varphi) r^{n-1} e^{-\zeta r}$$

Equation 31. STO equation. N : normalization constant; $Y_{l,m}$: spherical harmonic functions that depend on the angular momentum quantum number l and the magnetic quantum number m_l ; n : principle quantum number; ζ : constant related to the effective charge of the nucleus.

However, it is relevant to note that STO basis sets lack an adequate description of orthogonal orbitals with different n . Moreover, STOs are highly computationally expensive when applied to large molecules. As a result, GTOs have been developed to facilitate the computation of *ab initio* calculations in large systems.

$$\chi_{GTO} = NY_{l,m}(\theta, \varphi) r^{2n-2-l} e^{-\zeta r^2}$$

Equation 32. GTO equation. N : normalization constant; $Y_{l,m}$: spherical harmonic functions that depend on the angular momentum quantum number l and the magnetic quantum number m_l ; n : principle quantum number; ζ : constant related to the effective charge of the nucleus.

GTOs are characterized by their mathematical simplicity, which makes them computationally less demanding. The reason is that the product of two Gaussian functions at different centres is equivalent to a single Gaussian function centred at a point between those two centres. As a result, a two-electron integral on three or four different atomic

centres can be reduced to an integral on only two centres, which is much more efficient from a computational standpoint.

Despite this, GTOs suffer from a significant disadvantage when it comes to describing electrons around the nucleus. For instance, the orbital 1s has a cusp at the atomic centre that is well described by STOs but cannot be accurately described by GTOs. Since GTOs do not accurately describe atomic orbitals near the nucleus, large basis sets are required to achieve a reasonable level of accuracy, which in turn requires more computational power.

In response to this problem, contracted Gaussian functions (cGTO) have emerged, combining STO and GTO's strengths. By combining a set of primitive GTOs, these functions can describe atomic orbitals more efficiently.

3.7.2. Contracted Basis Sets

The contracted Gaussian-type orbitals (cGTOs) alleviate the abovementioned problems by combining several GTOs. Over the years, several cGTOs have been developed, particularly those developed by J. Pople (197,198), R. Ahlrichs (199), and T. Dunning (200).

3.7.2.1. Pople style basis sets

The Pople's basis set family consists of some of the most frequently used basis sets. Pople's basis sets, combined with methods like the B3LYP DFT functional (described later in this chapter), have led to excellent results for a wide range of molecular systems (46,201–210). For this reason, it was one of the methods of choice used in this work.

It belongs to the family of split-valence basis sets since it is based on the fact that chemical bonding affects valence and core orbitals differently. As valence orbitals are widely affected, it makes sense to use a valence basis function that is more flexible than a core basis function

For example, the 6-31G basis set uses six primitive Gaussian functions (denoted by the "6") to describe the core atomic orbitals. On the other hand, the "31" after the hyphen means that the valence orbitals are composed of two basis functions each. The first is composed of a linear combination of three primitive Gaussian functions, while the other is

composed of only one primitive Gaussian function. In other words, 6-31G is a split-valence double-zeta basis set. The letter “G” at the end denotes that the primitive functions used are Gaussian functions.

a. Diffuse and polarization functions

Pople’s basis sets are usually employed in a more complete form to better represent intermolecular interactions, with the addition of polarization (211) and diffusion (212) functions. For example, the 6-31G(d,p) basis set contains one set of first polarization functions applied on all atoms. Additionally, diffuse functions can be included to describe the electron density of ions more accurately. Similarly to the polarization function, diffuse functions can be applied solely to heavy atoms (6-31+G) or to all atoms (6-31++G).

3.7.2.2. Ahlrichs basis sets

Another split-valence basis set family is the def2 basis sets of the Ahlrichs group. Compared to the Pople basis sets, these are more consistent across the periodic table, hence being the default recommendation in the ORCA software. Their naming scheme is def2-XVP, where X can be S, TZ, or QZ, representing double-zeta, triple-zeta, and quadruple-zeta, respectively.

Def2 basis sets already include polarization functions by default (hence the letter P on their name, e.g., def2-SVP). However, contrarily to the Pople’s basis sets, initially, they did not have any way to include diffuse functions. Nowadays, this matter has been solved with the development of new def2 basis sets that include diffuse basis functions, which can be distinguished by adding the letter “D” to the end of the basis set name (e.g., def2-SVPD). However, due to their increased computational cost, they should be used strictly when necessary.

Regarding their use, a double-zeta basis set like def2-SVP is usually enough for initial geometry optimizations, resulting in reasonably good structures. However, as a rule of thumb, further optimizations should be performed with a triple-zeta basis set like def2-TZVP. At this point, molecular geometries will be fairly converged.

3.7.2.3. Dunning's correlation consistent basis sets

Even though the mentioned approaches provide accurate results, particularly regarding the molecules' geometry, several other basis sets include a more accurate description of their electrons.

Currently, the cc-pVNZ basis set family is widely used to provide a more accurate description of the orbitals at an affordable computational cost. Developed initially by Dunning, these basis sets were improved to include polarization (represented by "p") and electronic correlation descriptions (represented by "c") with the "N" taking the form of D, T, Q, 5, 6, and so on, which corresponds to double-, triple-, quadruple-, quintuple-, sextuple-zeta, etc. (200).

Similarly to the other basis sets, diffuse functions can be added to cc-pVNZ functionals to improve the description of anions and long-range interactions. This results in the subfamily of basis sets known as aug-cc-pVNZ (213).

Unfortunately, cc-pVNZ basis sets are computationally more expensive than Pople's and Ahlrichs', which prevents their application to geometry optimization of large molecular systems. Therefore, it is common to perform an initial geometry optimization using a Pople's basis set, such as 6-31G(d,p), followed by an energy refinement through single-point calculations using cc-pVNZ basis sets.

a. Basis set extrapolation

In modern computational chemistry, quantum chemical calculations are performed using a finite set of basis functions. When the finite basis is expanded towards an (infinite) complete set of functions, calculations using such a set are said to approach the complete basis set (CBS) limit.

A basis set extrapolation is a method of estimating the properties of a system from a limited number of calculations. This is done by extrapolating the results of these calculations to a larger basis set.

The family of Dunning's correlation consistent basis sets was designed with the systematic convergence of post-Hartree–Fock calculations (described in detail in the following sections) to the complete basis set (CBS) limit using extrapolation techniques.

A typical workflow is to perform single-point energy calculations with a crescent series of basis sets (e.g., cc-pVDZ and cc-pVTZ) and then extrapolate the energies to the CBS limit. Currently, all this process is performed by the ORCA software in an automatized manner.

3.8. Electronic Hamiltonian

3.8.1. *Hartree-Fock*

An exact wave function can only be determined for systems with one electron, such as the hydrogen atom. An atom or molecule with more than one electron is called a many-electron system. Unfortunately, the Schrödinger equation cannot be solved exactly for those systems. The Hartree-Fock method, however, can be used to obtain an accurate wave function by taking an initial guess at the wave function. In this approach, a multi-electron wavefunction is written as a product of single-electron wavefunctions. In conceptual terms, we convert a multi-electron atom into a collection of individual electrons located in individual orbitals whose spatial characteristics and energies can be studied separately.

Using the Hartree-Fock method, single-electron wavefunctions, referred to as atomic (if pertaining to single atoms) or molecular orbitals, can be used to describe a many-electron system.

It all started in 1927 when Douglas R. Hartree introduced a procedure called Self-Consistent Field (SCF) (214). This method is a practical application of the variational principle, meaning that it will iteratively modify the wavefunction until a predetermined threshold is reached, resulting in an energy closer and closer to the actual energy of the ground state. To achieve this, a set of three-dimensional atomic nuclei coordinates is used to create an initial guess of molecular orbitals using the SCF method. Next, the Fock matrix is generated and diagonalized for SCF convergence testing. In the event the SCF test fails to meet the predefined thresholds, a new Fock matrix is generated and tested. Otherwise, the Fock matrix is used to calculate the properties of the system, and the calculations are complete. SCF cycles are repeated until the criteria have been met or a software-imposed maximum number of cycles has been reached.

3.8.1.1. Many-Electrons System

In a many-electron system, there are a large number of electrons that interact, repelling each other. Thus, if these electron repulsions are neglected, a many-electrons system can be treated as a sum of n one-electron equations. The wave function ($\Psi(r_1, \dots, r_N)$) can be approximated as a product of individual electron wavefunctions:

$$\Psi(r_1, \dots, r_N) \cong \psi_1(r) \psi_2(r) \dots \psi_N(r)$$

Equation 33. Wavefunction as a product of individual electron wavefunctions.

This approximation is known as a Hartree product. However, this represents a problem as electrons should be indistinguishable among themselves and therefore do not obey the principle of antisymmetry of the wave function. This was one of the points that led to the development of the Hartree-Fock theory as we know it.

3.8.1.2. Hartree-Fock Theory

Hartree-Fock was the first method to treat many-electron systems in computational chemistry successfully. In HF theory, the wavefunction is described as the product of all the orbitals.

However, the HF theory introduces an approximation regarding the interelectronic repulsion, where the movement of the electrons depends only on the average field of the other electrons and the electrostatic field created by the nuclei. Consequently, each electron's wave function can be expressed as a function of the system's electrostatic field generated by all other electrons.

Considering that each electron must obey Pauli's exclusion principle (215), every one-electron function ($\phi(r)$) must be multiplied by its spin function (α or β):

$$\chi(x_N) = \phi(x_N) \cdot \alpha$$

Equation 34. One-electron spin-orbital.

Consequently, the wave function (Ψ_{HF}) of such a system can then be written as a product of N spin-orbital functions ($\chi(x_N)$):

$$\Psi_{HF}(x_1, x_2, \dots, x_N) = \chi_1(x_1) \chi_2(x_2) \dots \chi_N(x_N); \quad x = x, y, z, \omega$$

Equation 35. Wavefunction as a product of N spin-orbital functions.

3.8.1.3. Pauli Exclusion Principle and Slater Determinants

The Pauli exclusion principle states that two electrons with the same spin cannot occupy the same orbital (215). As a result, the wave function must be antisymmetric to allow the interchange of any two electrons with different spin states in the same final orbital.

In order to fulfil such a requirement, an antisymmetric wave function can be constructed using one Slater determinant (216) (**Equation 36**). A Slater determinant is a mathematical expression that describes the wavefunction of a many-electron system. It is named after John Slater, who introduced it in 1929 to ensure the antisymmetry of a many-electron wavefunction and, consequently, the Pauli exclusion principle.

$$\Psi(x_1, x_2, \dots, x_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(x_1) & \chi_2(x_1) & \cdots & \chi_N(x_1) \\ \chi_1(x_2) & \chi_2(x_2) & \cdots & \chi_N(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(x_N) & \chi_2(x_N) & \cdots & \chi_N(x_N) \end{vmatrix}; \quad \mathbf{x} = x, y, z, \omega$$

Equation 36. Slater determinant for an N -electron system. $\chi(\mathbf{x})$: spin-orbital where $\mathbf{x}$ denotes the position and spin of a single electron.

The Slater determinant represents every possible combination for an N -electron system in which the columns correspond to each electron's spatial spin wave function and the rows to its coordinates.

3.8.1.4. Hartree-Fock Limit: Electron Correlation

The HF equations can be used with additional determinants to account for more electronic phenomena. By doing this, the wave function starts to converge toward an exact solution of the electronic Schrödinger equation.

The SCF approach provides a simple way for the HF method to test and achieve accurate molecular geometries and properties, but this is only true for “well-behaved” molecular systems. Due to this, HF can only be used successfully for closed-shell electronic configurations and systems with low degeneracy of molecular orbitals. Furthermore, HF treats electrons as non-interacting particles, which is not true since electrons repel each other by having the same charge. Therefore, there is an energy difference between the one obtained with HF and the exact solution of the non-relativistic Schrödinger equation – the electron correlation energy. This difference results from the overestimation of the electron repulsion in HF.

Closed-shell systems usually present good results with HF due to having a minimal correlation energy contribution. However, these correlation effects play a more significant role in chemical reactions involving open-shell systems or electron transfers, where electron correlation is essential for accurately describing the chemical system.

3.8.2. *Post-Hartree-Fock Methods*

Hartree-Fock (HF) was one of the first attempts to apply the Schrödinger equation to more complex systems containing more than one electron. However, due to its significant limitations, HF is unsuitable for accurately studying most chemical phenomena. Its main weakness is that it relies on the average contribution of electrons, its “mean-field”. As a result, interactions between electrons are not explicitly considered, which neglects the electron correlation effect. By incorporating electron correlations into Hartree-Fock methods, post-Hartree-Fock methods improve upon Hartree-Fock. Some of these methods include the Configuration Interaction (CI), the Multiconfiguration Self-Consistent Field (MCSCF), the Møller-Plesset (MP) Perturbation theory, and the Coupled Cluster (CC) theory. Nevertheless, despite these methods being more accurate, they are also more computationally demanding, which means they are not always suitable for the study of large molecules.

3.8.2.1. Configuration Interaction (CI)

In the HF method, a finite basis set expansion yields a finite set of molecular orbitals. As a result, a basis set with N_o members will have N_o molecular orbitals. Furthermore, when the N_o molecular orbitals are ordered from lowest to highest energy half will be doubly occupied with electrons, which corresponds to the ground-state wavefunction. However, half of the virtual unoccupied antibonding orbitals remain empty.

Singly excited states are created when a single electron is promoted from an occupied spin-orbital to an unoccupied virtual spin-orbital. Similarly, a double excited state can be achieved if a second electron follows the same behaviour but for a different unoccupied virtual spin-orbital. In other words, multiple excited states that respect the required electronic symmetry can be formed. In addition, these states, also called determinants, can be linearly combined to form a configuration state function (CSF).

Accordingly, we can write the exact ground-state wavefunction (Ψ) as a linear combination of all possible Slater determinants that arise from a complete set of spin orbitals:

$$\Psi = c_0 \Psi_0 + \sum_{a,p} c_a^p \Psi_a^p + \sum_{\substack{a < b \\ p < q}} c_{ab}^{pq} \Psi_{ab}^{pq} + \sum_{\substack{a < b < c \\ p < q < r}} c_{abc}^{pqr} \Psi_{abc}^{pqr} + \dots$$

Equation 37. Exact ground-state electronic wavefunction Ψ where the c s represent expansion coefficients and the limits in the summation indices ($a < b$, etc.) ensure that a given excited determinant only appears once in the summation.

The configuration interaction (CI) method is an ab initio method based on this approach.

When the exact ground-state wave function is computed using the CI method, the energy calculated is the exact non-relativistic ground-state energy. As a result, the correlation energy corresponds to the difference between this exact energy and the one computed with HF and an infinite basis set (i.e., at the HF limit).

The CSF provides an exact wave function for a particular system, but it is computationally impossible to solve it with an infinite set of Slater determinants constructed from an infinite set of spin-orbitals. However, when only the determinants of a particular symmetry are included in the calculation, this cumbersomeness can be alleviated. The full CI is calculated by considering all CSFs for a certain symmetry and a finite basis set. Given a specific basis set, the full CI is the best CI calculation that can be made.

Thus, the correlation energy can be calculated using the difference between HF SCF and full CI calculations using the same basis set for CI calculations. The resulting basis set correlation energy becomes closer to the exact value if the basis set is enlarged in both calculations.

Despite the promising results, CI's application is computationally too expensive for its application in the study of interesting chemical problems. As a result, truncated versions of CI methods have been developed to account only for the ground-state and doubly excited states (DCI) or the ground-state, singly excited, and doubly excited states (SDCI).

Moreover, CI calculations are also criticized for their lack of size consistency. In other words, the energies obtained by CI methods are not proportional to the number of electrons in the system. Thus, the total energy of two particles separated by an infinite distance does not equal the sum of their energies. Additionally, the magnitude of the size

consistency error increases with increasing molecular system sizes. On the positive side, the CI method obeys the variational principle.

3.8.2.2. Multiconfiguration Self-Consistent Field Methods

Unlike the CI method, in which the expansion coefficients c_X (**Equation 37**) are calculated with HF and kept throughout the entire calculation, the multiconfigurational self-consistent field (MCSCF), a refined variation of the CI method, reoptimizes the coefficients throughout the calculation. While MCSCF is computationally demanding due to these additional calculations, it yields more accurate results by including fewer configuration state functions (CSFs).

In particular, the complete active space self-consistent field (CASSCF) method is a method that uses the MCSCF approach to study excited states. This method has three types of spatial wave functions: inactive, virtual, and active.

Inactive orbitals are made up of wavefunctions with the lowest energy that are doubly occupied. On the other hand, virtual orbitals are composed of the highest energy wave functions, which are entirely unoccupied. Lastly, active orbitals include orbitals with intermediate energy that do not fall into either of the previous two categories. Thus, the CSFs considered by a CASSCF calculation only depend on the states of electrons in the considered active orbitals.

As a result, the successful application of CASSCF is highly dependent on the choice of active orbitals. Therefore, CSF choices are based on molecular orbital theory, considering all bonding, non-bonding, and antibonding orbitals as active orbitals.

Although MCSCF methods have improved accuracy by using fewer CSFs than CI methods, size consistency remains a problem.

3.8.2.3. Møller-Plesset Perturbation Theory

Since the CI method lacks size consistency, the Perturbation Theory (PT) was developed to provide a consistent approach to finding the correlation energy.

There are, however, disadvantages to the PT method, including the fact that it is not based on the variational principle. Thus, it is not guaranteed that the calculated energies will always be higher than the exact energies. In addition, since energies can be higher or

lower than the exact energy, one cannot ensure that the result evolves toward the exact solution's direction.

The Many-Body Perturbation Theory (MBPT) is the application of the PT method to a system with many interacting particles. This approach was first introduced by C. Møller and M.S. Plesset in the form of the Møller–Plesset Perturbation Theory (MPPT) (217).

MPPT uses the zero-order Hamiltonian (H_{HF}) from the HF-SCF method, which can be written as the sum of one-electron Fock operators (f_i):

$$H_{HF} = \sum_{i=1}^{N_e} f_i$$

Equation 38. Zero-order Hamiltonian (H_{HF}) given by the sum of N_e one-electron Fock operators (f_i), where N_e is the number of electrons in the system.

Calculation of the perturbation (H_{pert}) is based on the difference between the electronic Hamiltonian (H_{elec}) and the zero-order Hamiltonian (H_{HF}):

$$H_{pert} = H_{elec} - H_{HF}$$

Equation 39. Perturbation contribution. H_{elec} : electronic Hamiltonian.

As a result, for electron i , the perturbation ($H_{pert}(i)$) can be expressed as:

$$H_{pert}(i) = j_0 \sum_j \frac{1}{r_{ij}} - \sum_m \{2J_m(i) - K_m(i)\}$$

Equation 40. Perturbation contribution for electron i . i and j : electrons; m : occupied molecular orbitals; J_m : Coulomb operator; K_m : exchange operator.

Essentially, the MPPT method takes advantage of the HF method and attempts to include electron correlation contributions through perturbations to the Fock operators.

It is possible to apply perturbation equations to integer orders of zero to n , but MP0 and MP1 provide a null contribution to the energy calculation. Therefore, only for perturbation equations of higher order (MP2, MP3, MP4, etc.), thus the electron correlation contribution begins to be appropriately accounted for in the Hamiltonian. Unfortunately, MP calculations get more costly with the number of basis functions used, so some high-order MP methods are not practical for larger systems.

Even though MP methods are size consistent, they do not rely on the variational principle, so the effect of applying evermore complete basis sets can be unpredictable. Moreover, higher-order MP methods do not necessarily provide better results, even though the accounted electron correlation contribution also increases.

3.8.2.4. Coupled Cluster (CC)

Coupled-cluster theory (CC) is another alternative to CI and MPPT (218). Like MPPT, the CC method is size consistent but does not follow the variational principle.

By introducing the cluster operator (C), the CC method establishes a relationship between the exact electronic wave function (Ψ) and the HF wave function (Ψ_0):

$$\Psi = e^C \Psi_0$$

Equation 41. According to the CC method, the exact electronic wavefunction Ψ is a product of the exponential operator e^C , and the HF wavefunction Ψ_0 .

As defined by **Equation 42**, the exponential operator (e^C) is defined by a series expansion that represents the sum of the effects of one-electron, two-electron, three-electron, up to N-electron excitation operators (**Equation 43**).

$$e^C = 1 + C + \frac{1}{2!} C^2 + \frac{1}{3!} C^3 + \dots$$

Equation 42. Exponential operator e^C defined by a series expansion.

$$C = C_1 + C_2 + \dots + C_N$$

Equation 43. Cluster operator C .

Since the CC method is too computationally demanding, it is common for the cluster operator to be truncated. As a result, the method only includes double excitation determinants (CCD), single and double excitation determinants (CCSD), or single, double, and triple excitation determinants (CCSD(T)).

It is possible to perform more accurate calculations with the same computational effort as CI and MPPT due to the mathematics behind the cluster operator. The CC methods, in particular, allow the inclusion of triplet excitation determinants (CCSD(T)) with relatively low computational costs. Thus, CCSD(T) is, nowadays, considered the “gold standard” of

accurate computational calculations, with an error of less than 1 kcal.mol⁻¹. Thus, benchmark studies of other functionals and methods often use CCSD(T) as a reference.

a. Domain-Based Local Pair Natural Orbital (DLPNO) Coupled-Cluster Method

Even though CC methods, such as CCSD(T), provide a relatively good balance between accuracy and computational cost, they are still very demanding for systems larger than a few dozen atoms. Therefore, it is virtually forbidden to use CCSD(T) in the study of biomolecular systems.

A novel approach developed by Neese and his colleagues aiming to reduce the computational cost associated with using the CCSD(T) method has been developed, the Domain Based Local Pair Natural Orbital (DLPNO), which takes advantage of the locality of the electron correlation component of CCSD(T). DLPNO-CCSD(T) classifies electron pairs as weakly or strongly correlated according to a threshold energy value for pair correlation energy. CC is then used to compute strongly correlated electrons, while MP2 is used to compute weakly correlated electrons. Moreover, the DLPNO method is highly flexible, being possible to alter its accuracy by just changing a few controllable thresholds. It has been demonstrated that DLPNO-CCSD(T) is 99.9% accurate compared to the canonical CCSD(T) method, reaching differences below 1 kcal.mol⁻¹ while doing so at a much lower computational cost. Besides, the DLPNO-CCSD(T) method scales almost linearly with system size. For all these reasons, the DLPNO-CCSD(T) method has been used with success in several studies of enzymatic catalytic mechanisms, where more than 100 atoms were treated with QM (219–223).

3.8.3. *Density Functional Theory*

Density functional theory (DFT) has become one of the most popular computational methods for computing molecular electronic structures. By using DFT, we can express the energy of an electronic system in terms of its electron probability density, ρ . Therefore, for a Ne-electron system, $\rho(r)$ is a function that gives the total electron density at a particular point in space, r . As such, we can obtain a single energy value for any given function $\rho(r)$ by treating the electronic energy E as a functional of the electron density (denoted as $E[\rho(r)]$).

But first of all, what is a functional? In simple terms, a functional is a function of a function. Mathematically, a functional takes a function as its input and outputs a single number. For example, consider the functional $F[f(x)]$ that can be described as:

$$F[f(x)] = \int_{-1}^1 f(x) dx$$

Equation 44. Hypothetical functional $F[f(x)]$.

So, if hypothetically, $f(x) = x^2 + 1$, the functional of $f(x)$, $F[f(x)] = \frac{8}{3}$.

Moreover, a key advantage of DFT is that a N_e -electron molecule can be represented by a single three-dimensional function, the electron density, without being constrained by permutation symmetry. On the other hand, a wavefunction approach relies on a $3N_e$ -dimensional function with many constraints to ensure that it is fully antisymmetric. In addition, the wavefunction of a N_e -electron molecule has cusps where electrons approach each other, whereas the electron density has none because it does not explicitly consider inter-electron distances.

Thus, a system of N electrons can be seen as a continuous distribution of electrons influenced by the external potential of positively charged atomic nuclei. Therefore, electron density can be directly related to electrons' kinetic and potential energies.

The electron density at the nucleus region $\rho(r_A)$ of an atom A with an atomic charge Z_A is given by the following equation. Since nuclei are point charges, they can be regarded as local maxima in electron density $\rho(r)$.

$$\frac{\partial}{\partial r_A} \bar{\rho}(r_A)|_{r_A=0} = -2Z_A \rho(r_A)$$

Equation 45. Electron density at the nucleus. $\bar{\rho}$: spherically averaged density; Z : atomic number.

In short, DFT rests on two fundamental theorems proved by Kohn and Hohenberg and on a later derivation of a set of equations by Kohn and Sham that we will see in the following sections.

3.8.3.1. Hohenberg-Kohn Theorems

According to two fundamental theorems by Hohenberg and Kohn, DFT can be applied to quantum mechanics (224).

The first of the Hohenberg-Kohn theorems states that the ground-state energy of a many-electrons system, E_0 , is a unique functional of the ground-state electron density $\rho_0(r)$, which in turn is a function of only three spatial coordinates (x , y , and z).

$$E_0 = E [\rho_0(r)]$$

Equation 46. Ground-state energy E_0 as the functional of the electron density $\rho_0(r)$.

The second theorem states that the electron density that minimizes the energy of the overall functional is the true ground state electron density.

$$E [\rho(r)] \geq E [\rho_0(r)]$$

Equation 47. Second theorem of Hohenberg-Kohn.

This theorem can be seen as the DFT version of the variational principle. However, while in theory, DFT should be variational, the variational principle introduced by the second theorem of Hohenberg-Kohn only applies to the exact XC functional. Since the exact XC functional is unknown, approximate XC functionals are used, which means that the variational principle does not work in practice, as we can get energies lower than the ground-state energy. This is contrary to Hartree-Fock, which is purely variational (i.e., its energies are always higher than the actual value).

3.8.3.2. Kohn-Sham Equations

As of now, only the general formalism of DFT has been presented. However, no guidance has been provided regarding how to apply DFT in practice. It was not until 1965 that W. Kohn and L.J. Sham introduced the first approach that allowed us to do so (225,226). In order to construct a set of equations that could be solved, Kohn and Sham considered a hypothetical reference system consisting of N_e -non-interacting electrons that “felt” an arbitrary external potential $v_{ref}(r)$ so that the electron density of the reference system ($\rho_{ref}(r)$) would approximate the true electron density ($\rho(r)$). As such, the hamiltonian for the reference system is given by

$$h_{ref} = \sum_{i=1}^{N_e} h_i^{KS}; \quad h_i^{KS} = -\frac{\hbar^2}{2m_e} \nabla_i^2 + v_{ref}(r_i)$$

Equation 48. Hamiltonian for the reference system.

The one-electron Kohn-Sham orbitals ψ_m^{KS} are eigenfunctions of the one-electron Kohn-Sham hamiltonians h_i^{KS} :

$$h_i^{KS} \psi_m^{KS}(i) = \varepsilon_m^{KS} \psi_m^{KS}(i)$$

Equation 49. Kohn-Sham orbitals.

Despite the similarity between the equations for DFT and those for HF methods, the calculation of electron-electron interactions is slightly different. In contrast to HF methods that apply the Fock operator to all electron-electron pairs using two operators, DFT calculates the interaction using a one-body operator. The electron-electron interactions (Coulomb repulsion and exchange-correlation) are calculated in DFT from the interaction of each electron with the electron density. As a result, DFT calculations can be performed much faster than HF-based methods (for the same accuracy) since the number of integrals that need to be solved is reduced.

3.8.3.3. Exchange-Correlation Density Functionals

Due to the absence of consideration of electron interactions, **Equation 48** does not provide accurate results even though the DFT Hamiltonian is exact. Because of this, several approximate schemes have been developed to address this issue by adding exchange and correlation effects functions ($v_{XC}[\rho]$) to the Hamiltonian. DFT's implementation became almost semi-empirical due to these approximations, even though its formalism is not empirical.

Today, several functionals have been developed to minimize the primary source of error in DFT, i.e., the calculation of the exchange-correlation energy ($E_{XC}[\rho]$). The exchange-correlation energy is a sum of the contributions of both exchange energy ($E_X[\rho]$) and correlation energy ($E_C[\rho]$):

$$E_{XC}[\rho] = E_X[\rho] + E_C[\rho]$$

Equation 50. Exchange-correlation energy.

Due to its more significant contribution to energy, the exchange term usually prevails over the correlation.

A drawback of the DFT method is the need to use approximations to account for exchange-correlation effects. Although the variational method can still be applied, the

resulting energies will no longer be consistently higher or equal to the real system energy. As a result, the quality of the results no longer depends on how low the system's energy level is.

As mentioned above, many DFT functionals have been adjusted to reproduce accurate results. However, the adjustments included in the exchange-correlation functions are inadequate for developing a “perfect” functional. There are, however, some properties a “perfect” functional should fulfil:

1. In theory, self-interaction should not affect the energy functional. In other words, a one-electron system, such as the hydrogen atom, should have an exchange energy ($E_X[\rho]$) that cancels out the Coulomb energy and a correlation energy ($E_C[\rho]$) that is zero.
2. A linear scaling of the exchange energy ($E_X[\rho]$) is expected when electron coordinates are multiplied by a constant factor.
3. Unlike exchange energy, correlation energy ($E_C[\rho]$) is not governed by a direct scaling rule. However, scaling electron coordinates by a factor greater than one should increase the correlation (and vice versa).
4. For finite systems, the correlation energy ($E_X[\rho]$) approaches a negative constant as the scaling parameter increases to infinity.
5. Since the Coulomb potential corresponding to a single positive charge must be exactly cancelled for a neutral atom, the exchange potential ($E_X[\rho]$) should exhibit an asymptotic $-r^{-1}$ behaviour as $r \rightarrow \infty$.
6. Ideally, the correlation potential should exhibit asymptotic $-\frac{1}{2}\alpha r^{-4}$ behaviour, with α representing the polarizability of the $N_{elec} - 1$ system.

As a result, developing DFT functionals requires adjusting parameters to satisfy the above criteria or fit experimental data. Additionally, this approach leads to wide variation in the accuracy of DFT functionals for different molecular systems. Furthermore, unlike ab initio methods, where CCSD(T) is considered the “gold standard”, there is no “gold” DFT functional.

This will be followed by a discussion of some DFT functionals that have been developed and are generally employed by the scientific community. In general, methods are selected based on their popularity, determined by how many studies have used them with success.

The rationale is that a method that has been tried and tested repeatedly with success has proven its worth more than one that has not.

As opposed to *ab initio* methods, in which their quality is determined by how electron correlation is accounted for, DFT does not follow this principle because the exchange-correlation energy is mostly empirically determined. Nevertheless, Perdew's Jacob's ladder (Figure 13) is the most prominent attempt to categorize density functionals (227). Using electron density descriptors for the exchange-correlation term, functionals are ranked based on their degree of approximation, which can be roughly correlated with their expected accuracy. Therefore we will now look in more detail at some of these methods.

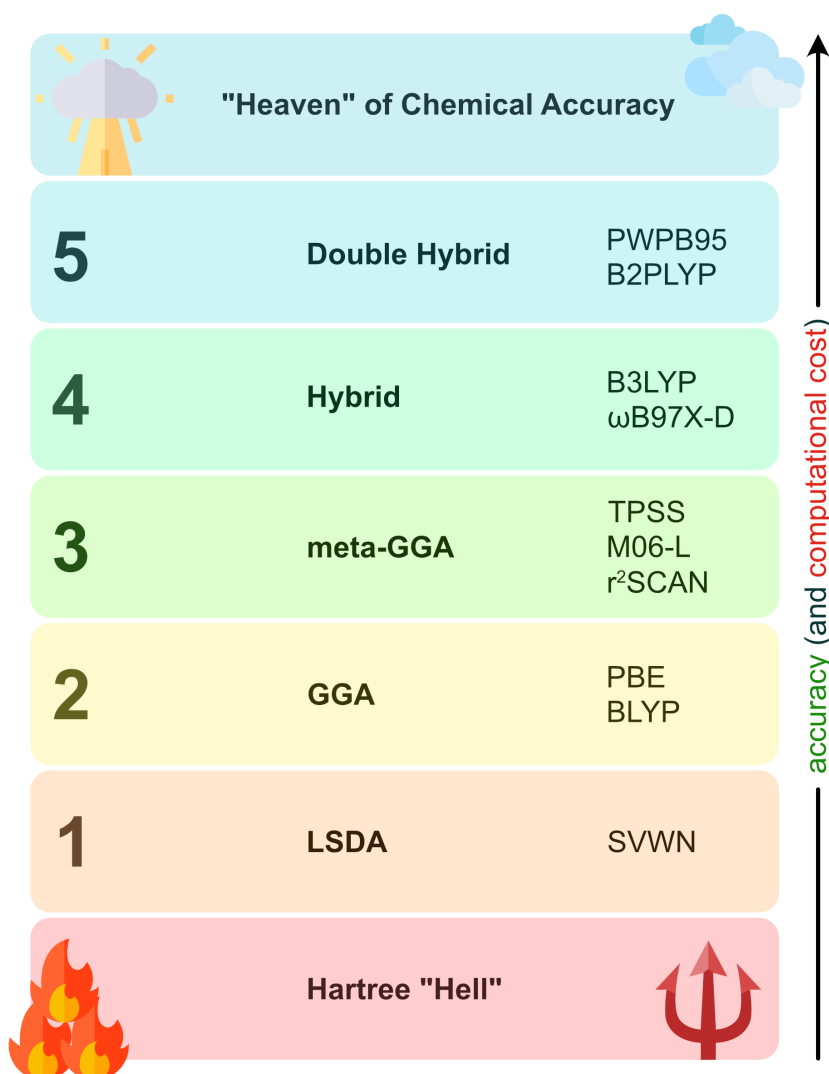


Figure 13. Categorization of the exchange-correlation density functionals according to Perdew's "Jacob's ladder". Adapted from (228).

a. Local Density Approximation (LDA)

In density functional theory (DFT), local density approximations (LDA) represent a class of approximations to the exchange-correlation (XC) energy functional based on the electronic density at every point in space (instead of derivatives of the density or the Kohn–Sham orbitals, for example). Homogeneous electron gas (HEG) models provide the best local approximations of XC energy, which can be applied successfully to real systems such as extended metallic systems (but not to typical molecular systems).

However, local-density approximations are essential to construct more sophisticated approximations to the exchange-correlation energy, such as generalized gradient approximations (GGA) or hybrid functionals. An important property in any approximate exchange-correlation functional is that it reproduces the exact results of the HEG for non-varying densities. Therefore, LDAs are often explicitly included within such functionals. Moreover, if the electron's spin is considered, the functional is called Local Spin Density Approximation (LSDA) instead.

b. Generalized Gradient Approximation (GGA)

The main drawback of LDA is that it assumes that density is constant everywhere, which results in an overestimation of correlation energy and an underestimation of exchange energy. To account for the inhomogeneity of the electron density, it is common to expand in terms of the gradient of the density. As a result, corrections can be made based on changes in density away from the coordinate, which makes the so-called generalized gradient approximation (GGA) more accurate than the LDA. By using GGAs, it is possible to obtain very reliable results for molecular geometries and ground-state energies. Some of the most prominent GGAs are the Perdew-Burke-Ernzerhof (PBE) exchange and correlation functionals (229,230) and the BLYP, formed by the Becke exchange (B88) (231) combined with the Lee-Yang-Parr (LYP) correlation functional (232).

Meta-GGAs, a natural development after GGA functionals, are potentially even more accurate. This is because the exchange-correlation potential of a GGA includes only the density and its first derivative. In contrast, the meta-GGA includes additional local information, such as the kinetic energy density or the Laplacian (second derivative) of the electron density. Popular meta-GGAs are the Tao-Perdew-Staroverov-Scuseria (TPSS) functional (233), the Minnesota functionals M06-L (234), M11-L (235), and MN12-L (236) by Truhlar and his coworkers, and the r^2 SCAN (237,238).

c. Hybrid Functionals

A hybrid functional is created by combining exchange and correlation GGA functionals with an HF exchange contribution (also known as exact exchange). Typically, some parameters are also included to define each term's weight in the hybrid functional. These parameters are usually determined by fitting the functional's predictions to experimental data.

For example, in the case of the popular B3LYP functional (239,240), its exchange-correlation energy is given by:

$$E_{XC}^{B3LYP} = [A \cdot E_X^{HF} + (1 - A) \cdot E_X^{LSDA} + B \cdot \Delta E_X^{B88}] + [C \cdot E_C^{LYP} + (1 - C) \cdot E_C^{VWN}]$$

Equation 51. Exchange-correlation energy of the B3LYP functional. A , B , and C are empirical parameters.

Where B3LYP empirical parameters A , B , and C are 0.20, 0.72, and 0.81, respectively.

In general, adding HF exchange to the functional significantly improves the accuracy of DFT results. Consequently, the amount of HF in the functional profoundly impacts the accuracy for different chemical systems and for calculating different properties. Therefore, it is crucial to make a careful choice to ensure that the DFT functional used accurately describes the chemical system of interest.

In summary, DFT is an excellent tool for studying various interesting chemical processes involving hundreds of atoms. Furthermore, the computational cost of DFT calculations is typically lower than wavefunction-based methods within the same accuracy range.

When studying reaction mechanisms, as was done in this work, it is common to use a hybrid DFT functional to optimize the molecular geometry of the system of interest and then use highly accurate wavefunction-based methods, such as CCSD(T) and DLPNO-CCSD(T), to refine the energy values of the resulting structures.

d. Composite Methods

Composite methods combine a well-balanced basis set with some corrections for the London dispersion, using the D3 (241) or D4 (242) models (described in more detail in the following section) and the inter- and intra-molecular basis set superposition error (BSSE), using the geometrical counter-poise scheme (gCP) (243–245). These methods allow us to

obtain an accurate molecular geometry on par with some hybrid functionals, so in standard thermochemical studies, it is rarely required to go beyond this level for structure optimizations. Currently, several composite methods are available, with B97-3c (246), PBEh-3c (245), and r^2 SCAN-3c (247), from Grimme's group being some of the most commonly used.

As such, r^2 SCAN-3c (247) was used in some of this work as the functional of choice for molecular geometry optimization. r^2 SCAN-3c is based on the meta-GGA r^2 SCAN functional that was combined with a tailor-made triple-zeta basis set (mTZVPP) as well as the D4 correction to London-dispersion and geometrical counter-poise (gCP) correction scheme to account for the basis set superposition error (BSSE).

Where methods like r^2 SCAN-3c shine is when they are applied in the context of a multi-level approach. Using r^2 SCAN-3c for structure optimization, we can obtain very accurate molecular geometries at a small fraction of the computational cost of more expensive functionals, like hybrid functionals. Then, we can simply perform a single-point energy calculation at a higher theoretical level, usually employing DLPNO-CCSD(T) with extrapolation to a complete basis set (CBS), to obtain our final energy values. The underlying idea of this workflow is that although r^2 SCAN-3c is not especially good with energies, it can provide structures on par with much more expensive functionals at a fraction of the computational cost.

3.8.3.4. Dispersion effects

A critical and evident error in DFT approximations is the absence of long-range correlations responsible for London dispersion. However, despite DFT's fundamental shortcomings, this can be easily fixed by including one of the several proven dispersion corrections, like D3 (241) or D4 (242), with little to no extra computational cost. Therefore, it may be argued that this is an essential aspect of any DFT treatment nowadays, and there is no chemical context where it should not be considered. Nevertheless, while dispersion effects are crucial to calculated energies, their influence on electron density appears negligible.

However, typical atom pair-wise additive approaches, like D3, are not self-consistent and thus do not correct the molecular orbitals nor any of the computed properties besides the energies. Therefore, methods like VV10 (248), a van der Waals correlation functional, have been developed that use electron density to account for the dispersion effects.

3.8.4. *Semi-empirical Methods*

There are still many challenges faced by theoretical chemists when it comes to describing large molecular systems accurately. Despite the ability to routinely predict the structure and energy of a few hundred atoms with Kohn-Sham density functional theory (DFT), it is still challenging to investigate the reactional space of the active site of an enzyme, for example.

Due to these limitations, semiempirical quantum mechanical (SQM) methods were developed to allow calculating large numbers of atoms in a reasonable amount of time. A valence-only minimal basis self-consistent field method derived from either the Hartree-Fock theory or the Kohn-Sham DFT is at the foundation of SQM methods. Due to drastic integral approximations, SQM methods are at least two orders of magnitude faster than *ab initio* QM methods. However, these methods have a significant reduction in accuracy and robustness. AM1 (249), PM3 (250,251), and PM6 (252) are some of the most widely used SQM methods.

Density functional tight binding (DFTB) methods have been gaining popularity over the past 20 years. Here, the Kohn-Sham DFT energy is expanded in terms of density fluctuations relative to atomic reference densities superimposed on each other. DFTB3 is the highest-level variant, which employs a self-consistent charge treatment, including density fluctuation terms up to the third order, and has been parametrized for many different elements (253).

GFN-xTB, a variant of DFTB3, which follows a global and element-specific parameter-only strategy, has been presented recently (254). The method has been used to calculate molecular geometries, vibrational frequencies, and noncovalent interaction energies. However, although GFN-xTB could identify the appropriate structures for less polar molecules, it did not describe polar and strongly hydrogen-bonded systems like sugars well, which motivated the creation of GFN2-xTB.

GFN2-xTB is the first parametrized tight-binding method that includes electrostatic and exchange-correlation Hamiltonian terms beyond monopole approximations (255). In addition, the partial atomic charge-dependent D4 London dispersion model is included. Like its predecessor, the parameters were fitted to produce reasonable structures, vibrational frequencies, and noncovalent interactions for molecules across the periodic table. Generally, this method focuses on organic, organometallic, and biochemical systems with a few thousand atoms or more. This study used GFN2-xTB as the first line of action to quickly explore every catalytic mechanism's reactional space.

3.8.5. Potential Energy Surface (PES)

A molecular system's potential energy surface (PES) describes the variation in total energy resulting from minor structural changes. A potential energy surface is a mathematical relationship that connects a molecular structure to its energy. In the case of a diatomic molecule, the potential energy can be represented by a two-dimensional plot with the distance between the nuclei on the x-axis and the energy at any bond distance on the y-axis. Nevertheless, for larger molecules, the potential energy surface has as many dimensions as the degrees of freedom within the molecule. This is because every point on the potential energy surface corresponds to a different energy value, determined by the molecule's bond distances, angles, and dihedrals. Therefore, a PES is usually represented as the illustration below, considering only two degrees of freedom of the molecule, with the energy plotted above the plane defined by them (**Figure 14**).

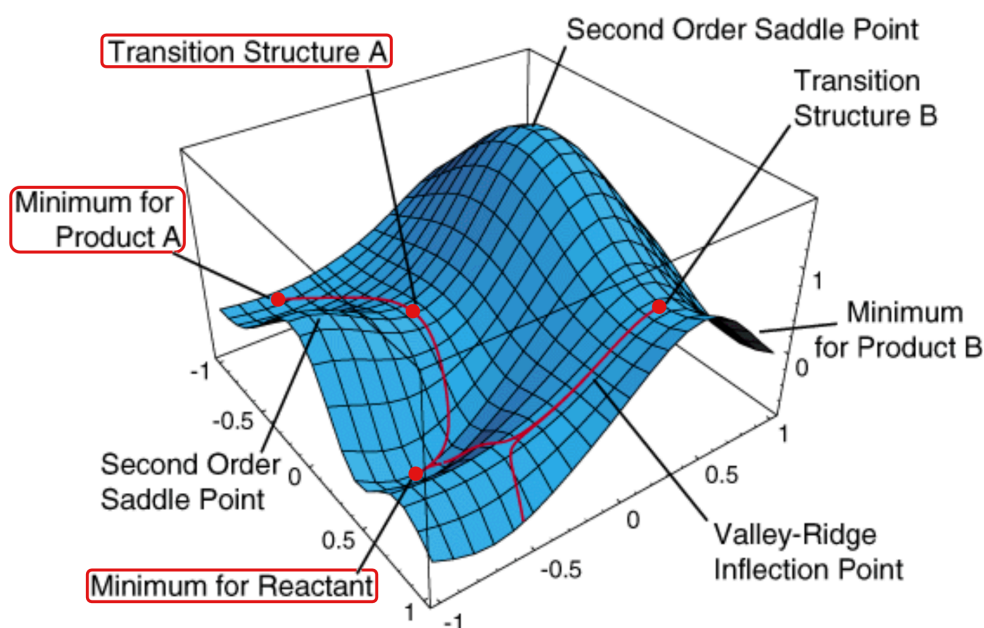


Figure 14. Model of a 3D potential energy surface showcasing, among other things, minima (a reactant and product), a transition state structure, and a reaction path. Adapted from (256).

There are three minima on this potential surface. A minimum is the bottom of a valley on the PES, so motion in any direction leads to higher energy. In a potential surface, a minimum can be a local minimum, the lowest energy point in a limited region, or a global minimum, the lowest energy point throughout the entire surface. The minima of a reaction path represent the equilibrium structures of a system, with different minima representing the products and reactants of a reaction path.

In contrast, peaks and ridges correspond to maxima along the potential energy surface. A peak is a maximum in all directions (i.e., along and across the ridge). A low point along a ridge corresponds to a local minimum in one direction (i.e., along the ridge) and a local maximum in the perpendicular direction. When a point is maximum in one direction and minimum in all others, it is referred to as a saddle point (**Figure 15**). Chemically, a saddle point is a transition structure connecting two equilibrium structures.

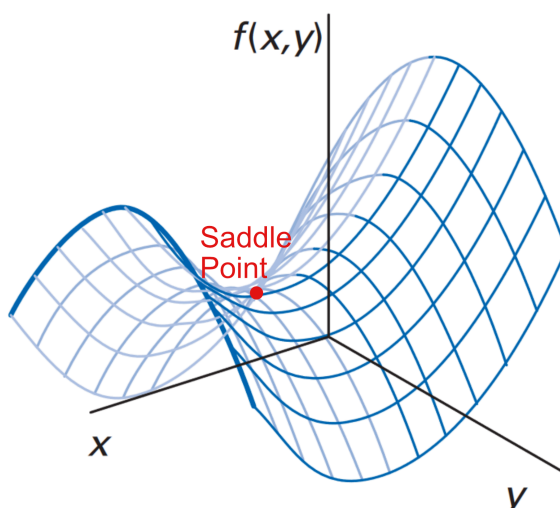


Figure 15. Saddle point at the surface of the saddle-shaped function: $f(x, y) = \frac{1}{2}k(x^2 - y^2)$. Adapted from (115).

3.8.5.1. Locating Stationary Points on the PES

Geometry optimizations can predict molecule equilibrium structures by locating minima on the potential energy surface. In both minima and saddle points, the first derivative of the energy, or gradient, is zero, so the forces on the nuclei are also zero (since gradients are the negative of forces). A point on the potential energy surface where the forces are zero is called a stationary point. Successful optimizations always find a stationary point, even if it is not the one intended. In geometry optimizations, a specific molecular structure is set as a starting point, and the potential energy surface is followed until a stationary state is reached (a minimum in the case of a reactant or product).

a. Locating Transition Structures

In addition to locating the ground state structures, optimizations can also locate transition structures since both correspond to stationary points on the potential energy surface. There are two main methods for locating a transition structure:

- Specifying a reasonable guess for the transition state geometry (such as the highest point of a scan, which will be discussed soon) and instructing the optimizer to locate a first-order saddle point.
- The second approach is called Nudged Elastic Band (NEB). In this approach, the reactant and product are used to find the minimal energy path (MEP) connecting both minima of the reaction. Then we can optimize the transition state using the MEP's highest energy point, as described earlier.

3.8.5.2. Characterizing Stationary Points

It has been noted that geometry optimization leads to a structure on the potential energy surface where the forces on the system are essentially zero. This structure may represent a saddle point or a minimum on the potential energy surface. A first-order saddle point is a transition state structure that links two minima by a maximum in one direction and a minimum in all other orthogonal directions.

Characterizing a stationary point requires two pieces of information:

- The number of imaginary frequencies.
- The normal mode corresponding to the imaginary frequency.

An n th-order saddle point is defined as a structure with n imaginary frequencies. Since ordinary transition structures are first-order saddle points, they have only one imaginary frequency. This happens because the force constant, k , (i.e., a second derivative with respect to displacement coordinates) for one normal mode is negative at the transition state, and since the square root of a negative number is an imaginary number, its frequency, ν , given by $\nu = \frac{1}{2\pi} \sqrt{\frac{k}{\mu}}$ (where μ is the reduced mass), is imaginary.

However, having a single imaginary frequency does not necessarily indicate that the transition structure is correct. In the potential energy surface, saddle points always connect two minima, but those minima may not correspond to the reactants and products of interest. When a structure yields an imaginary frequency, it indicates that the system

has a geometric distortion whose energy is lower than it is at the current structure (indicating a more stable structure). One must determine what the nature of this deformation is in order to fully understand a saddle point.

By visualizing the animation of the imaginary frequency on a molecular visualization software like VMD, one can determine whether the displacements correspond to the reaction coordinate of interest (**Figure 16**). Performing an intrinsic reaction coordinate (IRC) calculation to follow the reaction path and therefore determine the reactants and products explicitly is another, more accurate method of determining what reactants and products the transition structure connects.

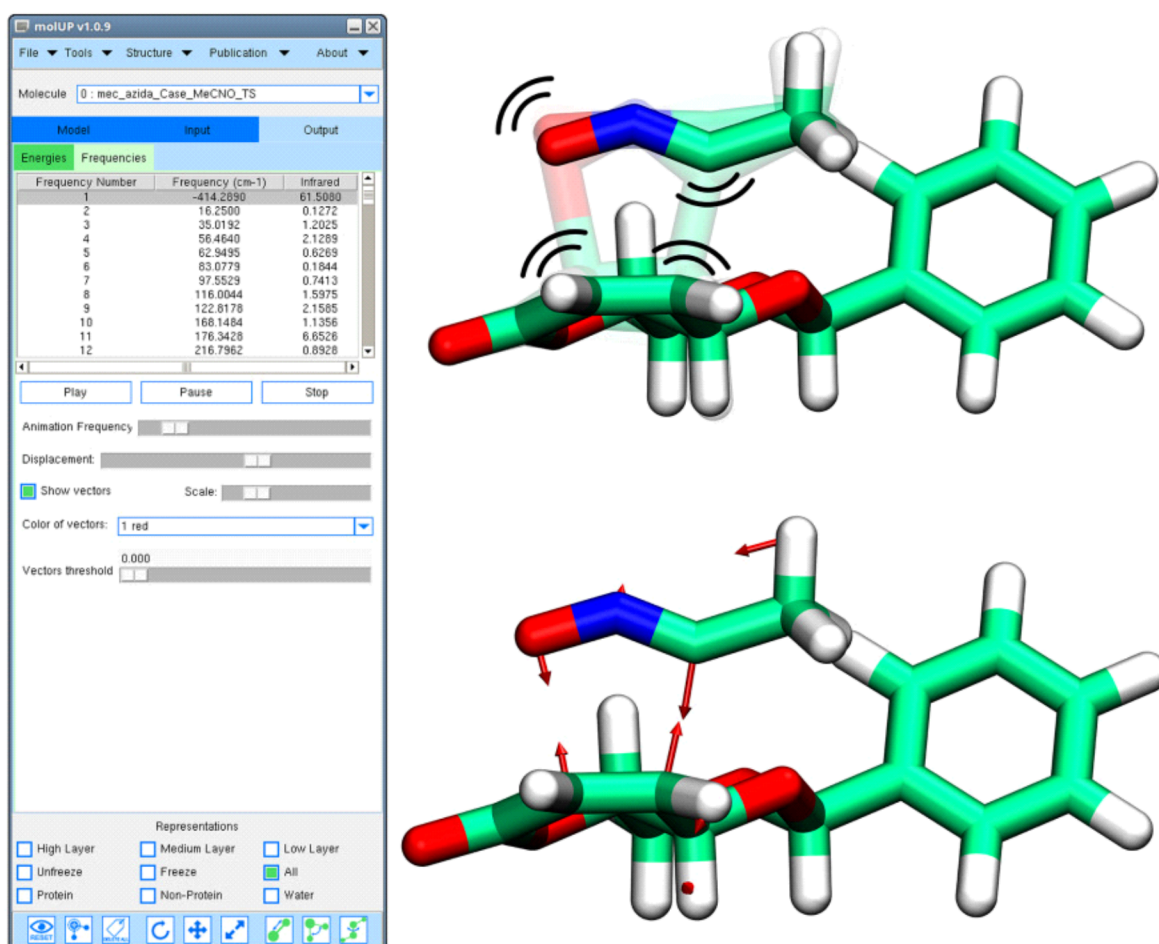


Figure 16. Analysis and visualization of the vibrational frequencies associated with a TS performed with molUP, a VMD plugin. If the structure has only one imaginary frequency (verified on the panel shown on the left), and the atom displacements observed correspond to the reaction coordinate of interest (represented by the red vectors), we can validate the structure as a real TS. Adapted from the original molUP article (257).

3.8.5.3. Potential Energy Surface Scans

Until now, we have focused on molecular geometries such as reactants, products, and transition structures in order to study reactivity. However, how do we find these stationary points on a large potential energy surface (PES)? We must first examine the actual landscape of this surface to figure out how the various stationary points are connected. Despite this, it is impossible to calculate or visualize the total PES for any but the simplest reactions. Therefore, the first step in studying any reaction step is to choose the reaction coordinates we intend to analyze carefully.

Using sophisticated energy derivatives, geometry optimization algorithms can find points of interest (minima and maxima corresponding to equilibrium and transition structures) on the PES. The PES can also be sampled in a region that corresponds to the process you are interested in. This type of calculation is referred to as a scan. The most common approach is to perform a relaxed scan, which does a partial optimization at each scan point, freezing the scan coordinate while optimizing all others.

Despite providing valuable insights into the structure of the PES, scan calculations do not define the lowest energy path between the two structures. Rather than simply stepping across the PES, we must follow the transition structure downhill to the reactants and products to plot the lowest energy pathway. The intrinsic reaction coordinate (IRC) calculation accomplishes this very purpose. This way, an IRC calculation can definitively link two minima on the potential energy surface through a transition state, verifying that a given transition structure does indeed connect the starting and ending structures we believe it connects. When this fact is confirmed, one can determine the activation energy of the reaction by comparing the reactants' and transition state's energy values.

3.9. Nuclear Hamiltonian

As we discussed in previous sections, QM methods mainly concern the calculation of the electronic energy. As a result of the Born-Oppenheimer application, the system's total energy can be divided into electronic and non-electronic terms (internuclear repulsion and nuclei kinetics) (**Equation 27**). However, this can be true only if the adjustment of electrons to nuclei movements is considered instantaneous. As already mentioned, this approximation is reasonably accepted because the mass of electrons is considerably lower than the mass of nuclei, and electronic phenomena play a significant role in chemical reactions. According to the Born-Oppenheimer approximation, the total energy

of a particular system requires the nuclear Hamiltonian (**Equation 27**) to be calculated and added to the electronic Hamiltonian.

3.9.1. Zero-point Energy Calculation

In order to determine the nuclear Hamiltonian contribution, further calculations are usually required. Since a molecule vibrates even at 0 K, the internal energy of a system (E_0) is calculated by adding the electronic energy (E_{elec}) and the zero-point energy (ZPE) correction (258):

$$E_0 = E_{elec} + ZPE$$

Equation 52. Zero-point energy (ZPE) correction.

3.9.2. Translational, rotational, and vibrational approximations

Moreover, thermal corrections are also computed to allow the computed energies to be compared with experimental results. As a result of thermal corrections, three terms associated with nuclei translation, rotation, and vibration can be taken into account in order to calculate the energy (E_T) for a given temperature T :

$$E_T = E_0 + [E_{trans} + E_{rot} + E_{vib}]_T$$

Equation 53. Energy at temperature T .

Approximated models are typically used to calculate those three terms. For example, rotational component calculations are done using the rigid rotor model, translational component calculations are done using the particle in a box model, and vibrational component assessments are made using harmonic oscillators.

As previously mentioned, in the study of a catalytic mechanism, each step is defined by three stationary points, a reactant, a transition state (TS), and a product. Hence, frequency calculations are needed to confirm that the structures obtained are stationary points. In the case of the TS, the vibrational calculation should include only one imaginary frequency corresponding to the vibration of the atoms involved in that reaction, whereas for the reactant and product, no imaginary frequency should exist.

Based on the harmonic oscillator, vibrational frequencies (ν) are calculated as follows:

$$\nu = \frac{1}{2\pi} \sqrt{\frac{k}{\mu}}$$

Equation 54. Frequency of a harmonic oscillator. μ : reduced mass; k : force constant.

This means that frequency calculation is only based on the force constants (k) associated with the displacement of the atoms. Using the second derivative of the energy with respect to the atomic coordinates, harmonic force constants are calculated. A saddle point (TS) structure may evolve in two opposite directions to decrease its energy (direction of reactants or products). Conversely, the movement across orthogonal directions increases the energy in both directions. According to **Equation 54**, this scenario results in an imaginary frequency since the second derivative (force constant) is negative. Therefore, a truly optimized minimum should have no imaginary frequencies, whereas the TS should only have one, as already mentioned.

3.9.3. Gibbs Free Energy

Thermal correction calculations include the calculation of enthalpy, entropy, and Gibbs free energy (259). Enthalpy is calculated as the expansive work within an ideal gas model at a given temperature. Thus, enthalpy (H) is expressed as:

$$H = E_T + k_B T$$

Equation 55. Enthalpy contribution estimated by thermal corrections. k_B : Boltzmann's constant; T : temperature.

The entropy of a nucleus is a sum of all entropic contributions from translational, rotational, and vibrational components, as well as the electronic entropy:

$$S = S_{trans} + S_{rot} + S_{vib} + S_{elec}$$

Equation 56. Entropy contribution estimated by thermal corrections.

However, a non-degenerated system is assumed to have zero electronic entropy (S_{elec}), and only its translation, rotation, and vibration movements affect its entropy.

Finally, in order to calculate Gibbs free energy for a given temperature (T), we use the following equation:

$$G = H - TS$$

Equation 57. Gibbs free energy.

In conclusion, nuclear hamiltonian calculations are fundamental to studying enzymatic mechanisms, enabling us to refine the final energies and determine the activation and reaction Gibbs free energies for each step of a catalytic mechanism. Calculating these quantities accurately is imperative for obtaining a reasonable estimate of the enzyme catalytic activity that can be compared to experimental results.

This page was intentionally left blank.

4. Hybrid Methods

Even though DFT represents a significant advance in applying QM calculations to a wide range of systems, it is still limited to a few hundred atoms. The method is suitable for a variety of chemical systems, but it cannot be applied to enzymes or other chemical systems with thousands or millions of atoms. Thus, additional approximations must be considered to study enzymatic mechanisms using computational methods.

A common and effective solution to this problem is the use of hybrid methods. Hybrid methods work by splitting a molecular system into two or more parts and treating those parts using different methods. The most popular hybrid method is QM/MM, where the reaction site is accurately treated with a QM method, like DFT, while the remaining part is treated with MM. This principle, however, can also be applied to other schemes, such as QM/QM or QM/QM/MM. Warshel, Levitt, and Karplus were awarded the Nobel Prize in Chemistry in 2013 for the development of this method (260–263).

Ultimately, hybrid methods enabled, for the first time, the application of computational methods to large and complex molecular systems, such as enzymes.

4.1. Quantum Mechanics/Molecular Mechanics

QM/MM is one of the most popular hybrid methods, dividing the system into two layers (**Figure 17**). The QM layer is usually comprised of a few hundred atoms and is treated using either wavefunction or DFT methods. On the other hand, the remaining atoms surrounding the QM layer are treated at a lower theoretical level, typically MM, which can handle thousands of atoms quickly.

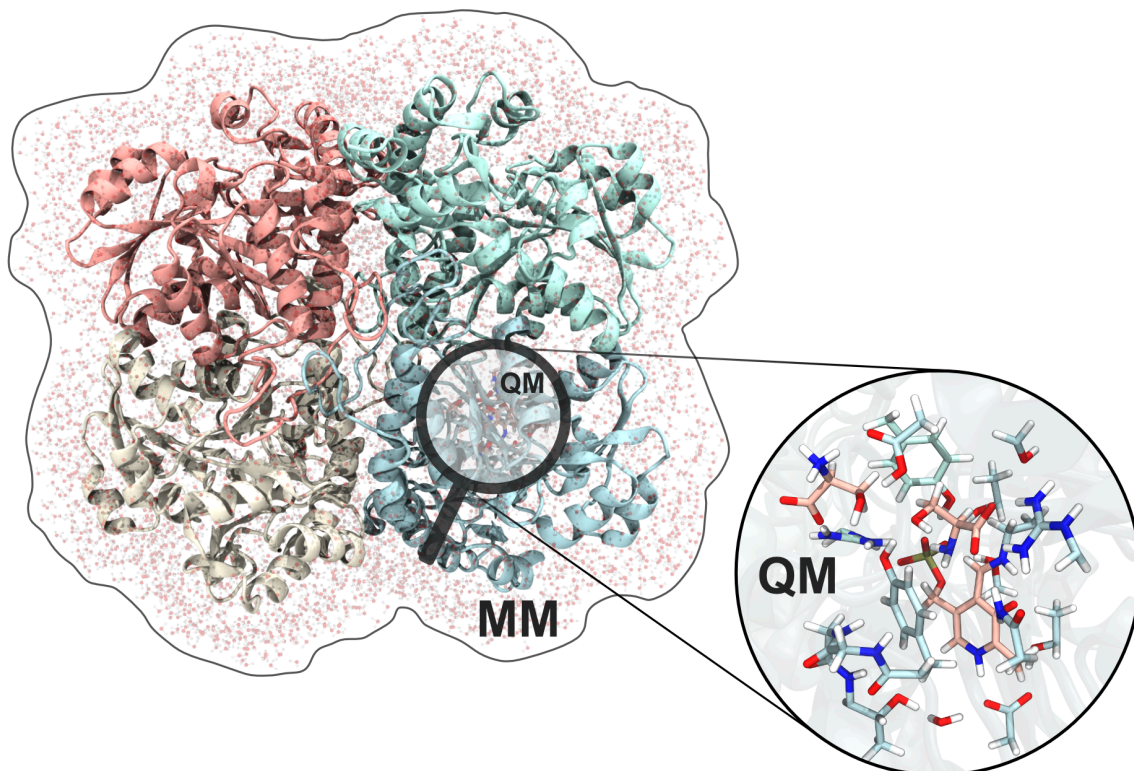


Figure 17. Representation of a QM/MM model. All the surrounded area (protein + water molecules) is treated with a MM method, while the residues within the highlighted area are treated with a QM method.

Despite the simplicity of such methods, there are some important considerations, particularly the contributions of each Hamiltonian and the boundary region.

4.1.1. Additive and Subtractive Methods

An easy method would be to calculate the total Hamiltonian by adding the Hamiltonians of each layer and the boundary regions between each pair of layers. This approach is called an additive method, and its energy equation ($E_{QM/MM}$) takes the following form:

$$E_{QM/MM} = E_{QM} + E_{MM} + E_{QM-MM}^{Interaction}$$

Equation 58. Energy function for the QM/MM additive method. E_{QM} : energy of the part of the system calculated with QM; E_{MM} : energy of the part of the system calculated with MM; $E_{QM/MM}^{Interaction}$: interaction energy between the QM and MM parts.

A challenge for the additive method, however, was calculating the contribution of the boundary region that intersects between the two layers ($E_{QM/MM}^{Interaction}$), which includes both

the bonded interactions (i.e., stretching, bending, and torsional contributions) as well as the nonbonded interactions (i.e., van der Waals and electrostatic interactions).

Thus, Maseras and Morokuma developed, in 1995, a subtractive method called ONIOM (Our Own N-layered Integrated molecular Orbital and molecular Mechanics) (264). The energy of a system computed using the QM/MM ONIOM approach (**Equation 59**) is calculated as the sum of the QM energy of the model system (E_{QM}^{Model}) plus the MM energy of the whole system (E_{MM}^{All}) minus the MM energy of the model system (E_{MM}^{Model}).

$$E_{QM/MM} = E_{QM}^{Model} + E_{MM}^{All} - E_{MM}^{Model}$$

Equation 59. Energy function for the QM/MM subtractive method. E_{QM}^{Model} : energy of the model part calculated with QM; E_{MM}^{All} : energy of the entire system calculated with MM; E_{MM}^{Model} : energy of the model part calculated with MM.

Furthermore, a multilayered approach (i.e., three or more layers) may be used, with each layer being treated with a different theoretical method.

Despite the ONIOM method's simplicity, it does come with a disadvantage - the interaction between the two layers is handled using MM. However, further improvements have been made to improve this aspect, as seen in the following sections.

4.1.2. Mechanical and Electrostatic Embedding

Two approaches have been developed to account for the contribution of atomic charges and geometries between layers during a QM/MM calculation. In the ideal case, all electrons would be included in the QM calculations. However, this would still pose the same problem since the main limitation of QM methods is the model size. For this reason, mechanical and electrostatic embedding schemes have been developed to mediate layer interactions.

4.1.2.1. Mechanical Embedding

The mechanical embedding approach is the simplest and most efficient in terms of computational calculation since the QM region does not feel the MM atoms during calculations. Instead, the surrounding MM atoms provide geometric constraints that limit

the possible movement of the atoms within the model system. As a result, it is as if the model system has been equilibrated in a vacuum.

The mechanical embedding approach involves optimizing the surrounding system with MM in each optimization step and the model system with QM in the gas phase. In other words, each layer is optimized by considering the forces of the adjacent layer, but no electronic interaction is taken into account.

Nevertheless, mechanical embedding can be applied successfully despite its simplicity if the model system is large enough to encompass all essential interactions.

4.1.2.2. Electrostatic Embedding

By incorporating partial charges from the MM region into the QM part of the calculation, electrostatic embedding provides a more accurate description of the electrostatic interaction between the QM and MM regions. However, including such charges in the QM Hamiltonian makes the calculation more time-consuming than using mechanical embedding. Moreover, the use of electrostatic embedding leads to some over-polarization of the model system since the charges of the surrounding system are kept immutable during the simulation, leading to artefacts when they are included in the QM Hamiltonian.

4.1.3. *Boundary Conditions*

Various methods have been discussed so far for coupling the calculation of different layers in a QM/MM system. However, those approaches only account for nonbonded interactions, which is not an issue if no covalent bond crosses between adjacent layers. However, in studying an enzyme's mechanism, the enzyme is too big to consider as a whole. Consequently, some covalent bonds will inevitably be split when constructing the QM/MM system.

The construction of a QM/MM system must, however, follow some rules regarding the truncation of covalent bonds:

- Cut only single bonds.
- It is preferred to cut nonpolar carbon-carbon bonds, although nonpolar single bonds may also be considered.
- Cyclic structures should not be cut.

- The MM region should be at least three bonds away from any bond breaking/formation.

However, by cutting a chemical bond, we are altering the valence of the atoms. Therefore, the link atom approach is used to address this question. In this approach, a link atom is placed in the direction of the truncated covalent bond, allowing the completion of the atomic valence of the involved atoms. Usually, a hydrogen atom is used because it is monovalent, small, and easy to compute. Therefore, the model system consists of the high-layer atoms (i.e., the region of the enzyme computed with QM) plus the hydrogen link atoms.

This page was intentionally left blank.

5. Enzymatic Kinetics

Several methods have been presented and discussed in this chapter concerning their application to studying chemical reactions, particularly enzyme mechanisms. As previously mentioned, DFT usually provides accurate geometries for the reactant, transition state, and product of a chemical reaction. After that, we can refine their energies using more accurate methods, such as DLPNO-CCSD(T), allowing us to calculate Gibbs free energies of activation and reaction that are more precise. So, why is the accuracy of the energy calculation so important?

The Gibbs free energies of activation ($\Delta G^\ddagger$) are essential checkpoints when studying enzymatic mechanisms since they provide a link between computational and experimental results.

Enzymatic kinetics states that a mechanism involving two or more steps is limited by the step with the slowest reaction rate. Thus, the slowest step has the highest Gibbs free energy of activation of the entire mechanism, representing the rate-limiting step.

The final energy profile of an enzymatic mechanism obtained by computational means includes the $\Delta G^\ddagger$ of each step (**Figure 18**). The step (or combination of sequential steps) with the highest barrier corresponds to the rate-limiting step. Therefore, we can corroborate our mechanistic proposal by comparing the calculated barrier for the rate-limiting step with the ones determined experimentally.

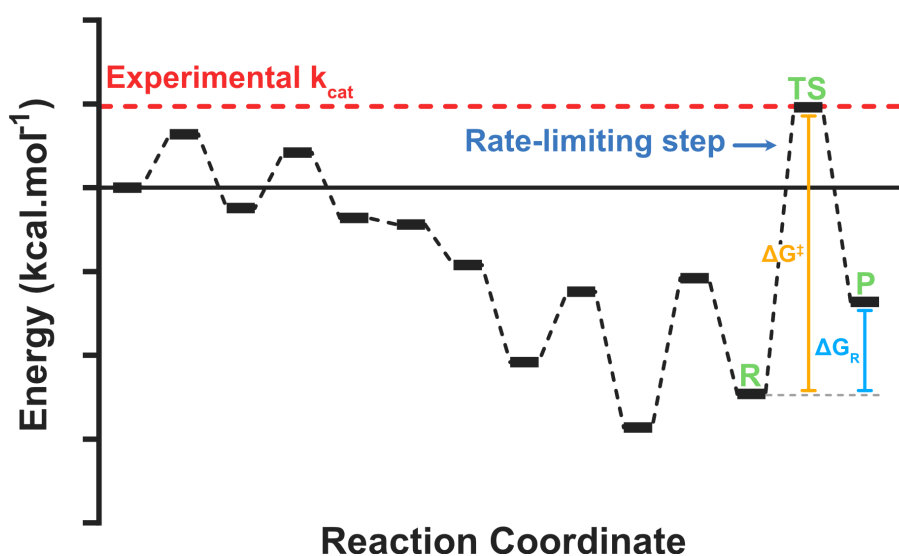


Figure 18. Schematic representation of a computationally determined energy profile. As observed the $\Delta G^\ddagger$ of the rate-limiting step (shown in orange) is identical to the energy barrier determined by the Eyring equation for the experimental k_{cat} (red). The Gibbs free energy of reaction, ΔG_R , of the rate-limiting step is also shown in blue.

Wet-lab scientists do not directly collect $\Delta G^\ddagger$ s. Reaction rates (k), turnover numbers (k_{cat}), and Michaelis constants (K_M) are collected instead. K_M is expressed in terms of molar concentration and is an indicator of the enzyme's affinity, which is seldom addressed in typical computational studies. On the other hand, k_{cat} describes how fast an enzyme is. It is primarily determined by the rate-limiting step of the mechanism and is expressed as the number of catalytic cycles per unit of time.

By establishing a relationship between k_{cat} and $\Delta G^\ddagger$, it would be possible to compare computed energies with experimental ones. As a result of Svante Arrhenius's discovery in 1889, that relationship was established and is known as the Arrhenius equation (265):

$$k = Ae^{-E_a/RT}$$

Equation 60. Arrhenius equation. k : reaction rate constant; A : Arrhenius constant; E_a : activation energy of the reaction; R : ideal gas constant; T : temperature.

Eyring then refined the empirical Arrhenius equation in 1935 to give it statistical mechanical support (266).

$$k = \frac{\kappa k_B T}{h} e^{-\Delta G^\ddagger/RT}$$

Equation 61. Eyring equation. k : reaction rate constant; κ : transmission coefficient; k_B : Boltzmann constant; T : temperature; h : Planck's constant; $\Delta G^\ddagger$: Gibbs free energy of activation; R : ideal gas constant.

With the Eyring equation, it is possible to determine a temperature-dependent relationship between the turnover number of a chemical reaction and the $\Delta G^\ddagger$ of the rate-limiting step. If the reaction evolves in the direction of the product without recrossing, the transmission coefficient (κ) usually equals one.

In summary, an experimentally determined enzymatic turnover number can be converted into the Gibbs free energy of activation for the mechanism's rate-limiting step. This value can then be compared to the $\Delta G^\ddagger$ calculated from the computational simulations to evaluate or refute a particular mechanistic proposition.

Furthermore, mutant forms of enzymes offer valuable clues about the importance of certain residues for catalysis or inhibition. In some instances, mutations can cause a significant decrease in the k_{cat} or even result in a loss of enzymatic activity. Consequently, this information indicates whether a specific amino acid residue is vital to the reaction.

Nonetheless, this is just one method of comparing computational data with experimental evidence. Some other approaches that could be combined include a comparison with intermediate structures (determined by X-ray crystallography or NMR), UV-Vis spectroscopy, and others.

This page was intentionally left blank.

C. RESULTS AND DISCUSSION

This page intentionally left blank.

This chapter was divided into four sections. **Sections C.1** and **C.2** focused on the study of two proteins, the enzymes Pdx2 and XometC. The goal was to characterize their complete catalytic mechanisms from an atomistic point of view. Hybrid QM/MM methodologies were used, resulting in Gibbs free energy values that provided important clues about how the catalytic process takes place and highlighted key steps, particularly the rate-limiting step that controls the overall kinetics of the catalytic process. The results agree very well with the available experimental data and provide a unified explanation of the experimental data regarding these enzymes. The resulting information from these studies was then distilled into the key aspects of each catalytic process, which can ultimately be used to develop new inhibitors (reversible or irreversible) targeting those enzymes.

Sections C.3 and **C.4**, on the other hand, are focused on the study of DNA, namely the dynamics and stability of 3D non-B DNA conformations, as well as DNA aptamers and their interaction with a potential target. These works resulted from collaborations, highlighting the increasing role of computational methods, such as molecular dynamics simulations, in this highly emerging field. A workflow was developed to consistently predict and simulate ssDNA molecules under physiological conditions. Moreover, DNA aptamers' affinity towards their target was predicted and ranked from best to worst, then confirmed experimentally.

Each section is composed of an initial summary of the work, followed by a complete portrayal of the resulting scientific article, accepted for publication in a research journal with a high impact factor in their respective field.

This page was intentionally left blank.

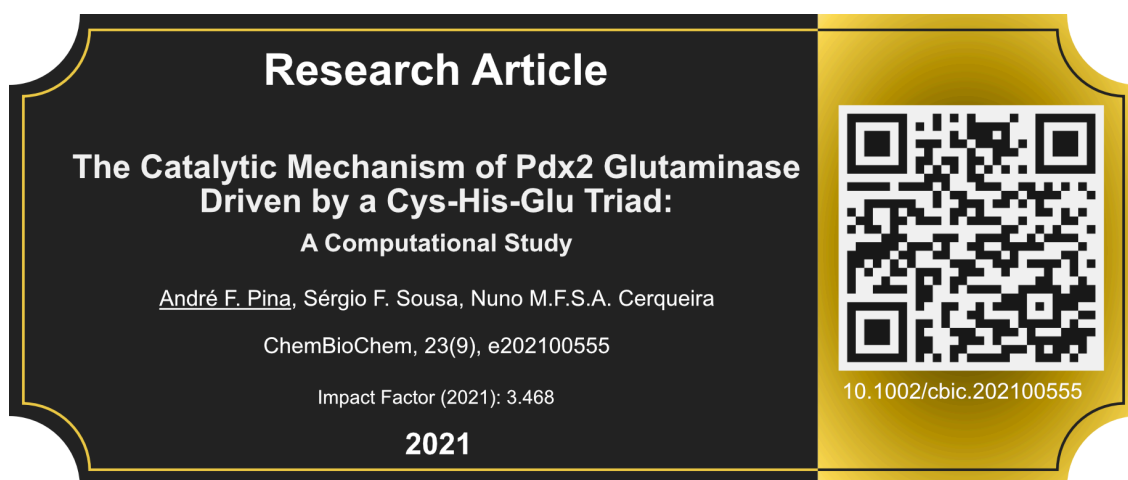
1. PLP biosynthesis

This page intentionally left blank.

1.1. Catalytic mechanism of the Pdx2 glutaminase subunit of PLP synthase

The catalytic mechanism of the Pdx2 glutaminase subunit of PLP synthase, which is essential to produce PLP in the Pdx1 synthase subunit, was studied in detail at an atomic level, employing the computational ONIOM hybrid QM/MM methodology.

André F. Pina performed all the calculations and wrote the article, which was revised with contributions from all authors. This work was published in the ChemBioChem journal, and the following content is almost a transcript of the published version. Additionally, this manuscript was selected as the cover feature of the issue in which it was published (**Appendix:** page 270).



Pdx2 employs a Cys-His-Glu catalytic triad to deaminate glutamine to glutamate and ammonia – the source of the nitrogen of pyridoxal 5'-phosphate (PLP). This enzyme is, therefore, a rate-limiting step in the PLP biosynthetic pathway of Malaria and Tuberculosis pathogens that rely on this mechanism to obtain PLP. For this reason, Pdx2 is considered a novel and promising drug target to treat these diseases. The results obtained show that the catalytic mechanism of Pdx2 occurs in six steps that can be divided into four stages: (i) activation of Cys₈₇, (ii) deamination of glutamine with the formation of the glutamyl-thioester intermediate, (iii) hydrolysis of the formed intermediate, and (iv) enzymatic turnover. The kinetic data available in the literature (19.1–19.5 kcal mol⁻¹) agree very well with the calculated Gibbs free energy barrier of the hydrolytic step (18.2 kcal.mol⁻¹), which is the rate-limiting step of the catalytic process when substrate is readily available in the active site. This catalytic mechanism differs from other known amidases in three main points: i) it requires the activation of the nucleophile Cys₈₇ to a thiolate; ii) the hydrolysis occurs in a single step and therefore does not require the formation of a second

tetrahedral reaction intermediate, as it is proposed, and iii) Glu₁₉₈ does not have a direct role in the catalytic process. Together, these results can be used for the synthesis of new transition state analogue inhibitors capable of inhibiting Pdx2, thus impairing diseases like Malaria and Tuberculosis.

The results of this work were presented at the following conferences:

1. (Poster) *BlmBS school 2019 – Bioinformatics meets BioSimulations in protein and DNA studies: from theory to practice*, **2019**, Lugano, Switzerland (**Appendix:** page 262)
2. (Poster) *VIII EJIBCE – Encontro De Jovens Investigadores Em Biologia Computacional Estrutural*, **2021**, Coimbra, Portugal (**Appendix:** page 265)
3. (Oral Communication) *Bioinformatics Open Days*, **2022**, Braga, Portugal
4. (Poster) 2022 IUBMB–FEBS–PABMB Young Scientists' Forum, **2022**, Lisboa, Portugal (**Appendix:** page 266)
5. (Poster) The Biochemistry Global Summit, **2022**, Lisboa, Portugal (**Appendix:** page 266)

1.1.1. Introduction

Tuberculosis and malaria are highly infectious diseases responsible for the annual death of millions of people worldwide (267,268). A significant factor hindering tuberculosis and malaria control is the high degree of resistance developed by their pathogens, *Mycobacterium tuberculosis* (*Mbt*) and *Plasmodium falciparum* (*Pf*), respectively, against the currently available drugs. For this reason, there is an urgent need for the identification of novel drug targets, as well as for the development of new anti-tuberculosis and anti-malarial drugs.

Although the survival and growth of *Mbt* and *Pf* pathogens rely on host-derived nutrients, some essential vitamins need to be produced on their own (10,40). Given this feature, the pathway of the biosynthesis of vitamin B₆ has been highly regarded as a potential drug target (10,12–14,16,269,270), giving new hope to possible treatments.

Vitamin B₆ is a collective term that refers to the vitamers pyridoxal, pyridoxine, and pyridoxamine, as well as to their respective phosphate esters. However, the term “vitamin B₆” is also used interchangeably with pyridoxal 5'-phosphate (PLP), its metabolically active form. PLP is an essential co-enzyme in many pathways, helping in the catalysis of several types of reactions, such as racemization, decarboxylation, beta-addition, beta-elimination, retro-aldol cleavage, transamination, alpha-elimination, among others (9,72,74–76,202,204,271,272).

Whereas humans, like other mammals, depend on vitamin B₆ uptake, previously referred pathogens, possess a *de novo* pathway for its biosynthesis, which is vital for their development and survival (10,28,59). This represents new hope in the development of new drugs against these diseases.

The *de novo* synthesis of PLP can occur by two different pathways: the deoxyxylulose 5-phosphate (DXP)-dependent and the DXP-independent pathways. While the DXP-dependent was the first to be discovered and is better characterized (22,24,25,29,273,274), the DXP-independent pathway is more ubiquitous among the three domains of life, being the one found in *Mbt* (10) and *Pf* (40). This pathway consists of a single enzymatic complex, the pyridoxal 5'-phosphate synthase (PLPS), a class I glutamine amidotransferase (GATase) composed of two distinct domains: Pdx2, the glutaminase domain, and Pdx1, the synthase domain (28,33). PLPS is a 24-mer enzyme composed of a dodecameric core of Pdx1 subunits of cylinder shape, to which twelve Pdx2 subunits bind, one to each Pdx1 subunit. Complex formation is required for the catalytic activity of either domain.

Pdx2, the glutaminase subunit, is responsible for ammonia production obtained by the hydrolysis of glutamine. The ammonia produced by Pdx2 is then channelled to the active site of Pdx1, where it is used as the source of the ring nitrogen of PLP (**Figure 19**).

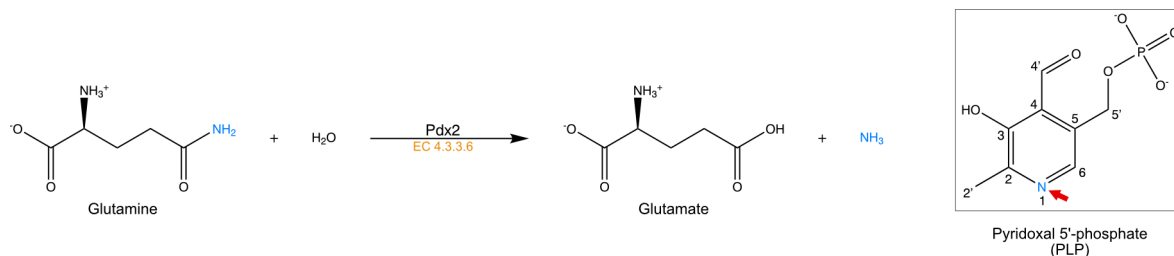


Figure 19. General reaction catalyzed by Pdx2. The ammonia produced by Pdx2 is the source of the ring nitrogen of PLP, denoted by a red arrow in the image.

The primary residues involved in the catalysis are Cys₈₇, His₁₉₆, and Glu₁₉₈ (40), which constitute the signature catalytic triad present in all class I GATases (275,276). However, there is some discussion regarding the role of Glu in the catalytic triad of class I GATases (275,277–281). It is uncertain if there is any proton transfer between the catalytic Glu and the His or if the role of glutamate consists only in correctly positioning the imidazole sidechain of histidine. Mutagenic studies in the glutaminase subunit of the *p*-aminobenzoate synthetase, a class I GATase, have shown that when glutamate is mutated to alanine or aspartate, the loss of activity is negligible when compared with a mutation in the cysteine or histidine residues, corroborating the hypothesis that glutamate is not crucial for the catalysis (277).

Another region of great importance in the active site of Pdx2 is the oxyanion hole, composed of the backbone nitrogen of Ala₈₈ and Gly₁₂₂. Its formation only occurs after the binding of the substrate to the Pdx1 subunit, preventing the unnecessary hydrolysis of glutamine by Pdx2. This indicates that the catalytic process can only occur when the assembly of the two subunits is complete.

Kinetic parameters of Pdx2 were determined experimentally, showing a reaction rate of 0.08 s⁻¹ (pH 8.0, 30 °C) for *P. berghei* (28), 0.11 s⁻¹ (pH 8.0, 30–37 °C) for *P. falciparum* (40,42), and 0.13 s⁻¹ (pH 8.0, 37 °C) for *B. subtilis* (33). According to the transition state theory (282) and using the Eyring equation (266), this experimental data indicates an activation barrier limit of 19.1–19.5 kcal mol⁻¹. In addition, mutagenic studies have shown that the catalytic His₁₉₆ (170 in *B. subtilis*) is essential for the catalytic process since when mutated for an asparagine, the glutaminase activity ceases (33).

In this work, the catalytic mechanism of the glutaminase domain of PLPS, Pdx2, from *Plasmodium falciparum* was studied using computational and theoretical means, namely Quantum Mechanics/Molecular Mechanics (QM/MM) methods. An atomistic and detailed description of the catalytic mechanism of Pdx2 will be provided, discussing the role of glutamate in catalysis, as well as the function of the conserved water molecule present in the active site.

1.1.2. Methodology

1.1.2.1. Model preparation

This study was conducted using the crystallographic structure of the chimeric Pyridoxal 5'-phosphate Synthase (PLPS) from *Plasmodium falciparum*, and *Plasmodium berghei* deposited on Protein Data Bank (PDB) with the reference 4ADS (28). The phosphate ions present in the active site of Pdx1 subunits, an artefact of the crystallization conditions, were removed. Additionally, the mutation His₁₉₆Asn was reverted in all Pdx2 subunits.

Subsequently, glutamine, the substrate of the Pdx2 enzyme, was introduced in its active site. This was accomplished by the superimposition of Pdx2 subunits of the 4ADS structure with the Pdx2 subunits of the 2NV2 from *Bacillus subtilis* containing L-glutamine in its active site (33). The model obtained can be visualized in **Figure 20**.

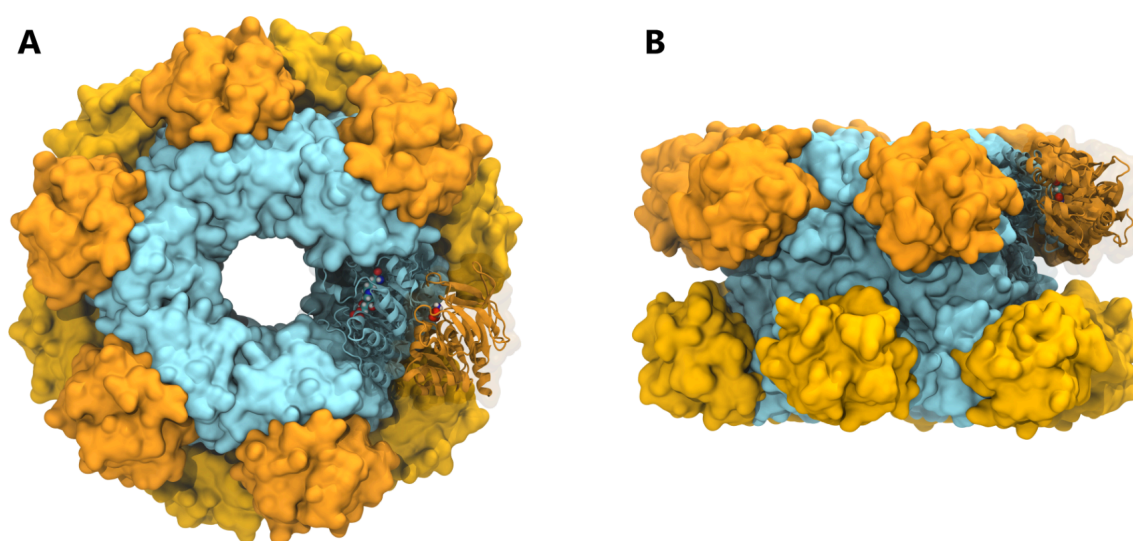


Figure 20. (A) Top and (B) Side view of the PLP synthase macrocomplex represented in QuickSurf. The Pdx1 and Pdx2 subunits are coloured in cyan and orange, respectively. One of the

Pdx1:Pdx2 dimers is represented in NewCartoon, with the substrates, ribose 5-phosphate and glutamine, represented in VDW.

1.1.2.2. Structure minimization

The system was prepared with the AmberTools software package, version 18 (283), using the ff14SB (129) and General Amber Force Field (GAFF) (132,133). For the natural substrate, L-glutamine, force field parameters were assigned with AMBER recommended protocol (using the Antechamber package) based on the GAFF parameters. RESP charges were derived with the Gaussian 09 (284) software using the Hartree-Fock method (HF/6-31G(d)). The protein was described by the AMBER ff14SB force field. The structure of the glutamyl-thioester intermediary, reactant in step 5 of the catalytic mechanism, was also parameterized with GAFF using the residuegen tool, part of the Antechamber package of AmberTools18. However, due to the absence of AMBER force field parameters for the thioester moiety of this intermediary, external parameters were added (285).

The protein and substrate were solvated with TIP3P (286) water molecules in a rectangular box so that each atom of the protein is distanced at least 12 Å from the edge of the periodic box. A physiological concentration of NaCl (150 mM) was introduced into the system. All residues were protonated by the LeaP package of AmberTools18 software (283) according to their physiological protonation state at pH 7.4. The PROPKA 3 software was used to predict the pK_a of the ionizable groups of the protein (120,121).

The minimization of the system was performed in four stages using AMBER18 (283). First, all water molecules were minimized (5000 steps), and cartesian positional restraints, with a restraint weight of 50 kcal.mol⁻¹.Å², were applied on all atoms except for the water molecules. Second, all hydrogen atoms were minimized (5000 steps), and the same restraint weight was applied on all non-hydrogen atoms. Third, all non-backbone atoms in the system were minimized (5000 steps), with the restraints applied on the protein backbone. Finally, a minimization of all the atoms in the system with no restraints was performed (100,000 steps).

1.1.2.3. QM/MM

The QM/MM model was built using the molUP (257) plugin for Visual Molecular Dynamics (VMD) software (287), obtained through the VMD Store (288). Due to the substantial size

of our system, only one Pdx2 subunit was included in the ONIOM QM/MM model. Therefore, the QM/MM model used in this study, built from the previously minimized structure, includes one Pdx2 monomer, the Pdx1 monomer that interacts directly with it, the glutamine in the Pdx2 active site, and all water molecules located at a maximum of 5 Å from the protein. In addition, all atoms that distance more than 15 Å from the active site were frozen during geometry optimizations.

The ONIOM QM/MM methodology was used in all the calculations, allowing us to split the model into different regions computed by different theoretical levels (264,289,290). The system was divided into two regions: the high-level (HL) layer, described with the density functional theory (DFT), computationally more demanding, and the low-level (LL) layer, described with molecular mechanics (MM), employing the same parameters used for the system minimization. The HL layer includes the glutamine, the catalytic triad Cys₈₇, His₁₉₆, and Glu₁₉₈, the backbone amide groups of Gly₁₂₂ and Ala₈₈ that form the oxyanion hole, and some other relevant backbone residues present in the Pdx2 active site, totalling 81 atoms (**Figure 21**). The remaining atoms were included in the LL layer. Hydrogen atoms were used as link atoms to complete the valence of the bonds crossing between the two layers of the ONIOM model.

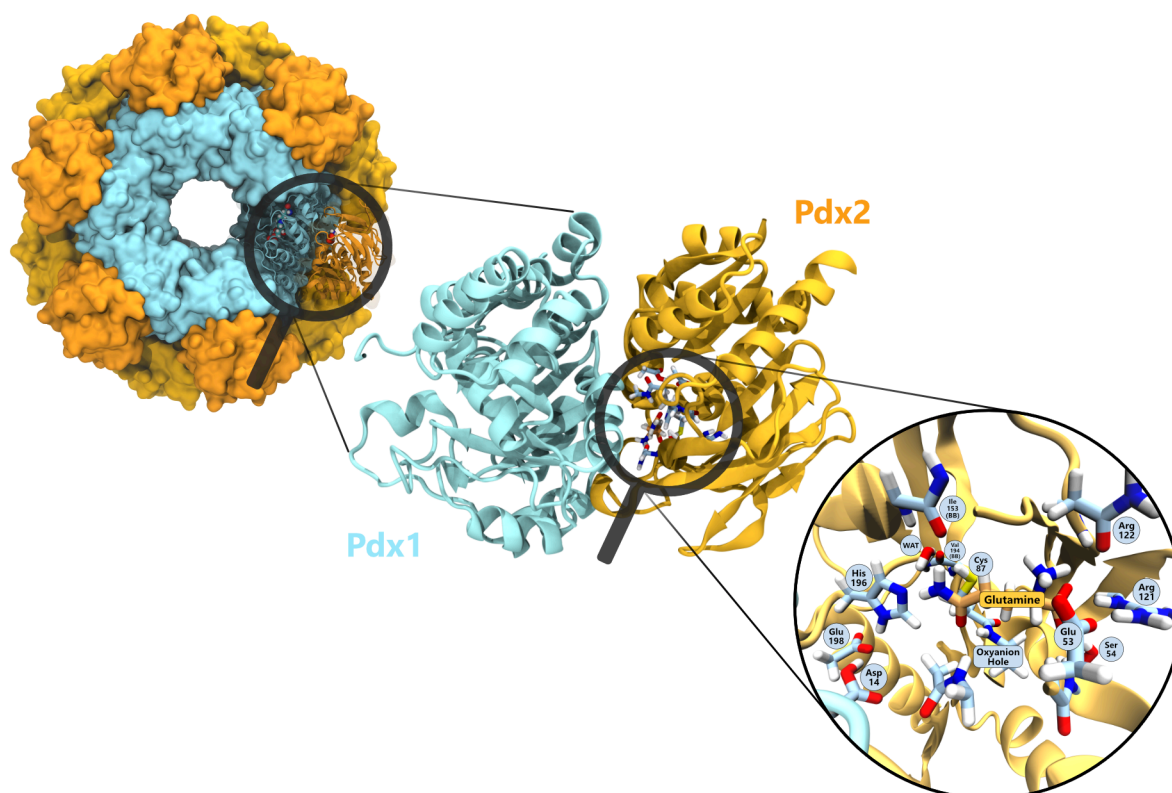


Figure 21. (Left) Structure of the dimer Pdx1:Pdx2 (cyan and orange, respectively) and Licorice representation of the HL layer. (Right) The substrate, glutamine, and active site residues are represented in Licorice and coloured in orange and blue, respectively.

The geometry optimization of the HL layer was carried out using the B3LYP (239) functional as available in Gaussian 09 (284). The basis set used was the 6-31G(d,p) from the Pople family (197,198,211,291,292).

The reaction space was explored through relaxed potential energy surface (PES) scans along the reaction coordinates implicated in each step of the reaction. The transition states (TS) were optimized using the structure with the highest energy obtained from the scans and further confirmed by the presence of only one imaginary frequency associated with the vibration along the reaction coordinate. The reaction and product of each step were obtained by the intrinsic reaction coordinate (IRC) method (293,294).

Vibrational frequency calculations were also conducted for the reactants and products at the same level of theory as the optimizations, and no imaginary frequencies were observed. The zero-point energy (ZPE), thermal, and entropic corrections were estimated at 298.15 K and 1.0 atm during the frequency calculation of TS and related minima structures.

Single-point energy calculations (SP) were performed using DLPNO-CCSD(T)/CBS for the HL layer as a way to refine the final energies (219,220,223). The Domain-based Local Pair Natural Orbital Coupled-Cluster method with Single, Double, and perturbative Triple excitations (DLPNO-CCSD(T)) was selected since it combines the high accuracy of the gold-standard method CCSD(T) while having a (low) computational cost similar to DFT (220–223). The SPs were carried out using the EXTRAPOLATE(3) keyword in the ORCA software version 4.2.1 (295,296). Three calculations using the basis set from the family cc-pVxZ (200,297) (where x is D, T, and then Q) were done, and the complete basis set (CBS) results were automatically extrapolated from that. The cc-pVQZ/C (298) auxiliary basis set, used for correlation fitting, was chosen. The final energy values obtained are a sum of the electronic energies of the HL layer resultant from the DLPNO-CCSD(T)/CBS calculation, the LL layer contributions previously calculated with the ff14SB force field, and the thermal corrections calculated with B3LYP/6-31G(d,p). This protocol has been used with success in the study of the catalytic mechanism of several enzymes (46,204–206,299).

The net atomic charges (NAC) and the bond orders calculated in each step of the catalytic mechanism were obtained by DDEC6 atomic population analysis, a refinement of the Density Derived Electrostatic and Chemical (DDEC) approach (300,301). This was accomplished through Chargemol version 3.5, a program written primarily for this purpose (302). In order to proceed to the analysis, the gbw file (GBW stands for “Geometry-Basis-Wavefunction”) obtained from the single-point energy calculation was converted to a wfx

file with the orca_2aim utility program part of ORCA software, allowing it to be read by Chargemol.

1.1.3. Results and Discussion

The proposals available in the literature about the catalytic mechanism of Pdx2 describe a sequential set of reactions culminating in the deamination of glutamine and formation of the glutamyl-thioester intermediate. Subsequently, the hydrolysis of this reaction intermediate occurs, and the product of the reaction, glutamate, can be released from the active site.

Based on the knowledge of other glutaminases (275,276), it has been proposed that the first step of the catalytic process starts with the nucleophilic attack of Cys₈₇ to the substrate. This would be concomitant with one proton transfer from Cys₈₇ to His₁₉₆ and another proton transfer from His₁₉₆ to Glu₁₉₈. However, all the attempts to model this reaction suggest that it is very unlikely to happen, being energetically unfavourable. Instead, the calculations demonstrated that Cys₈₇ needs to be first activated to a thiolate, and only then the nucleophilic attack of the substrate takes place.

Based on these findings, the complete catalytic mechanism of Pdx2 was divided into four main stages: i) activation of Cys₈₇, ii) formation of the glutamyl-thioester intermediate, iii) hydrolysis of the glutamyl-thioester intermediate, and iv) the enzymatic turnover, performing a total of six sequential steps. Each of these steps will be described and discussed in detail in the following sections.

1.1.3.1. Stage 1: Activation of Cys₈₇

Several crystallographic structures of different glutaminases (275,303–305) reveal the presence of a water molecule near the catalytic triad, as described by Bauer *et al.* (34). The same was verified in the active site of the structure used in this work. This indicates that such water molecule could directly participate in the activation process of the catalytic triad, behaving as a proton shuttle between Cys₈₇ and His₁₉₆. Based on this information, the first step was modelled in the presence of the water molecule, and two pathways were investigated: one where the water molecule directly participates in the activation process and another where it plays a secondary role.

The simulations were able to show that when the water molecule does not take part in the reaction, His₁₉₆ directly abstracts the proton from Cys₈₇ (**Figure 22**). In the reactants of this reaction, Cys₈₇ is protonated, and its proton is at 3.53 Å from nitrogen N ϵ^2 of His₁₉₆. In the transition state of this reaction (1013.4i cm⁻¹), the proton is no longer covalently bonded to Cys₈₇ (1.65 vs 1.36 Å in the reactants), becoming closer to N ϵ^2 of His₁₉₆ (1.34 vs 3.53 Å in the reactants) therefore favouring the proton transfer. In the reaction product, Cys₈₇ becomes anionic (charge of S_V is -0.71 vs -0.27 a. u. in the reactants), whereas His₁₉₆ becomes protonated and positively charged (charge of its imidazole ring is 0.70 vs 0.01 a. u. in the reactants). This reaction requires a Gibbs free energy of activation of 13.6 kcal.mol⁻¹, and it is endergonic in 11.7 kcal.mol⁻¹. The high reaction energy of 11.7 kcal.mol⁻¹ can be explained by the formation of two point charges in the Cys₈₇ and His₁₉₆.

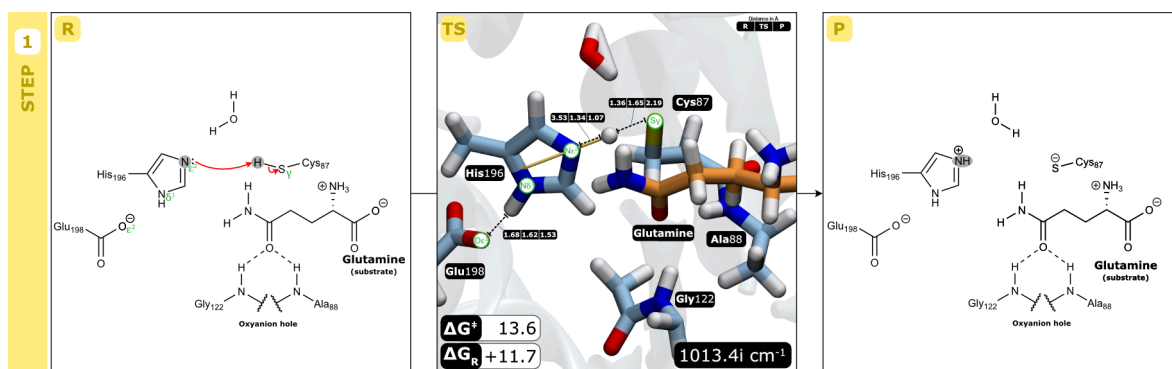


Figure 22. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 1. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. Glutamine, the substrate, is represented in orange, whereas the rest is in light blue. The yellow vectors illustrate the major vibrational modes corresponding to the imaginary frequency of the TS (shown in the bottom right corner). Relevant interatomic distances (in Å) for **R**, **TS**, **P** are shown, as well as the activation and reaction Gibbs free energies for this step.

On the other hand, the simulations were able to demonstrate that when the water molecule behaves as a proton shuttle between Cys₈₇ and His₁₉₆, the reaction is less favourable, requiring a higher Gibbs free energy of activation (16.9 kcal.mol⁻¹) and being even more endergonic (17.2 kcal.mol⁻¹) (**Supplementary Figure S1**).

These results demonstrated that even though the water molecule is positioned in a way where it could work as a proton shuttle, the activation of Cys₈₇ is favoured without its direct participation. However, the water molecule may have a crucial role in the correct alignment and orientation of the thiol group of Cys₈₇ towards the C₆ of the substrate.

This first step is required to improve the nucleophilicity of Cys₈₇, increasing its reactivity toward the substrate.

It is also worth mentioning that during either of the studied reactions, Glu₁₉₈ interacts with His₁₉₆ through one hydrogen bond, but Glu₁₉₈ never becomes protonated, not even at the end of the reaction, as it is commonly found in other catalytic triads (1.53 Å in the products vs 1.68 Å in the reactants). In the classical catalytic triad (Ser-His-Glu), Glu directly participates in the reaction by accepting the proton from His upon the nucleophilic attack of Ser to the substrate. However, our results show that throughout the entire catalytic process, Glu₁₉₈ is not protonated by His₁₉₆ or by any other residue. The distance between the O ϵ^2 of the Glu₁₉₈ and the H bonded to the N δ^1 of His₁₉₆ ranges between 1.51–1.78 Å, with an average of 1.64 Å, and its bond order ranges from 0.14 to 0.24, with an average of 0.19. These results are in agreement with the available experimental data, in which the mutation of Glu₁₉₈Ala does not affect the enzyme's catalytic efficiency (277). One possible explanation for this behaviour can be attributed to the bigger van der Waals radius of the sulfur atom (1.80 Å) of the thiolate of Cys₈₇, in comparison to the van der Waals radius of the oxygen atom (1.52 Å) of Ser. The closer contact with the residue His₁₉₆, even when Cys₈₇ is covalently bonded to the substrate, may compromise the proton transfer from His₁₉₆ to Glu₁₉₈.

1.1.3.2. Stage 2: Formation of the glutamyl-thioester intermediate

The next stage of the catalytic process involves the formation of the glutamyl-thioester intermediate, in which a reaction intermediate becomes covalently bonded to the enzyme through Cys₈₇. The results demonstrate that this stage occurs in three sequential steps (steps 2 to 4).

The first step of stage 2 (step 2; **Figure 23**) involves the nucleophilic attack of the anionic Cys₈₇ to the carbon C δ of the substrate (glutamine), leading to the formation of a tetrahedral intermediate that becomes covalently bonded to the enzyme.

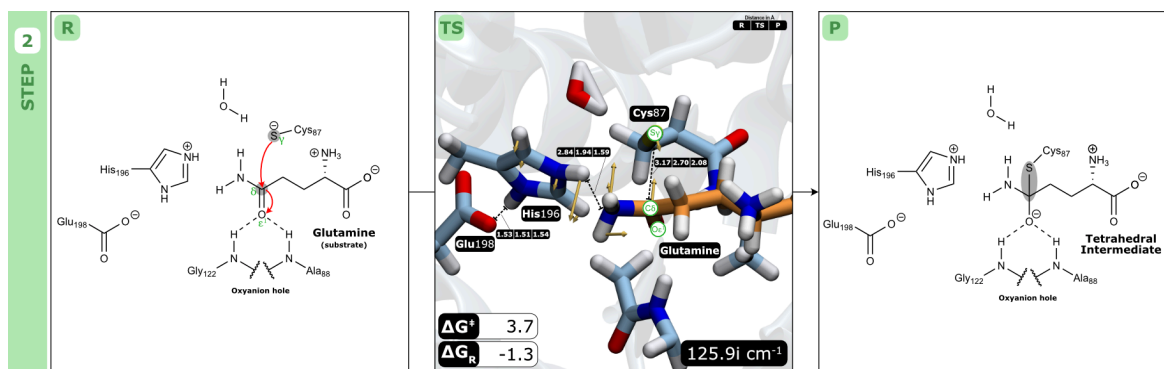


Figure 23. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 2. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. Glutamine, the substrate, is represented in orange, whereas the rest is in light blue. The yellow vectors illustrate the major vibrational modes corresponding to the imaginary frequency of the TS (shown in the bottom right corner). Relevant interatomic distances (in Å) for **R** **TS** **P** are shown, as well as the activation and reaction Gibbs free energies for this step.

In the reactants, the anionic Cys₈₇ (charge of S_γ is -0.71 a. u.) is 3.17 Å away from the carbon C_δ of the substrate. The positively charged His₁₉₆ (the charge of its imidazole ring is 0.69 a. u.) does not participate in this reaction, continuing to interact very closely with Glu₁₉₈ by one hydrogen bond (1.53 Å).

The transition state of this reaction is characterized by an imaginary frequency at $125.9i$ cm^{-1} . In this structure, the bond between the anionic Cys₈₇ and the substrate is starting to form (2.70 vs 3.17 Å in the reactants), as indicated by an increase of the bond order from 0.06 in the reactants to 0.15 in the TS. At the same time, a conformational rearrangement of the positively charged His₁₉₆ is observed, where its imidazole ring gets closer to the amino group of the substrate (1.94 vs 2.84 Å in the reactants).

In the reaction product, the substrate is now covalently attached to the enzyme by Cys₈₇ (2.08 vs 3.17 Å in the reactants), with a bond order of 0.52 (vs 0.06 in the reactants). At the same time, it is observed the elongation of the carbonyl group ($C_\delta-O_\epsilon$) of the substrate (1.29 vs 1.24 Å in the reactants), as well as a decrease in its bond order from 1.63 in the reactants to 1.47 . The charge of the tetrahedral intermediate becomes partially lodged in the oxygen of the carbonyl group (-0.77 vs -0.68 a. u. in the reactants), which is stabilized by two hydrogen bonds provided the backbone amide groups of Gly₁₂₂ (2.04 vs 2.05 Å in the reactants), and Ala₈₈ (1.87 vs 2.40 Å in the reactants), a region that is commonly known as oxanyon hole. The distance between the carbon C_δ and nitrogen N_ϵ of the intermediate also increased (1.51 vs 1.36 Å in the reactants), and its bond order

decreased (0.89 vs 1.30 in the reactants). His₁₉₆ remains positively charged (the charge of its imidazole ring is 0.66 a. u.), but it is now closer to the amino group of the formed tetrahedral intermediate (1.59 vs 2.84 Å in the reactants). Glu₁₉₈ continues to interact with His₁₉₆ by one hydrogen bond (1.54 Å).

The simulations indicate that the reaction requires a Gibbs free energy of activation of 3.7 kcal.mol⁻¹ and is exergonic in 1.3 kcal.mol⁻¹.

The catalytic process proceeds with the protonation of the amino group of the tetrahedral intermediate by His₁₉₆ (step 3; **Figure 24**).

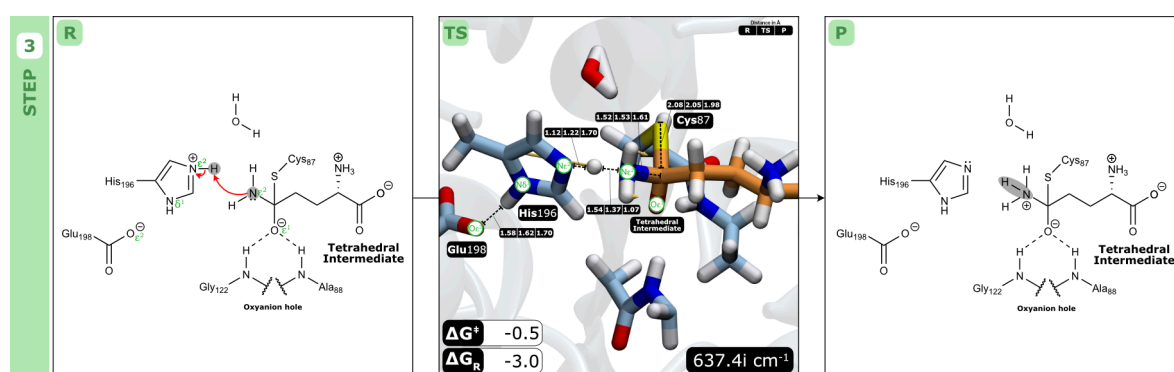


Figure 24. Structures of reactant (R), transition state (TS), and product (P) of step 3. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. Glutamine, the substrate, is represented in orange, whereas the rest is in light blue. The yellow vectors illustrate the major vibrational modes corresponding to the imaginary frequency of the TS (shown in the bottom right corner). Relevant interatomic distances (in Å) for R TS P are shown, as well as the activation and reaction Gibbs free energies for this step.

In the reactants, the amino group of the tetrahedral intermediate continues to be covalently bonded (1.52 Å; bond order: 0.88). Nonetheless, it is now very close to His₁₉₆ (1.54 Å).

This transition state is characterized by an imaginary frequency at 637.4i cm⁻¹. When compared with the reactant, the only significant change observed in this structure is that the proton is now equally shared between the nitrogen N_ε² of the tetrahedral intermediate and His₁₉₆.

In the reaction product, nitrogen N_ε² of the tetrahedral intermediate becomes protonated, and His₁₉₆ becomes neutral (0.04 vs 0.66 a. u. in the reactants). Consequently, the reaction intermediate becomes positively charged at the NH₃ substituent group (0.36 vs

-0.14 a. u. in the reactants). However, despite the formation of the ammonia group, the cleavage of the C_δ-NH₃ bond is not yet observed.

The results show that this reaction is very fast, requiring a Gibbs free energy of activation of -0.5 kcal.mol⁻¹ and being exergonic in 3.0 kcal.mol⁻¹. The low activation energy indicates that this step probably occurs spontaneously, and the minimum identified is most likely a short-lived specie, meaning that the reaction should proceed swiftly to the next step.

The formation of the glutamyl-thioester intermediate is only completed after the release of one ammonia molecule, which occurs in step 4 (**Figure 25**).

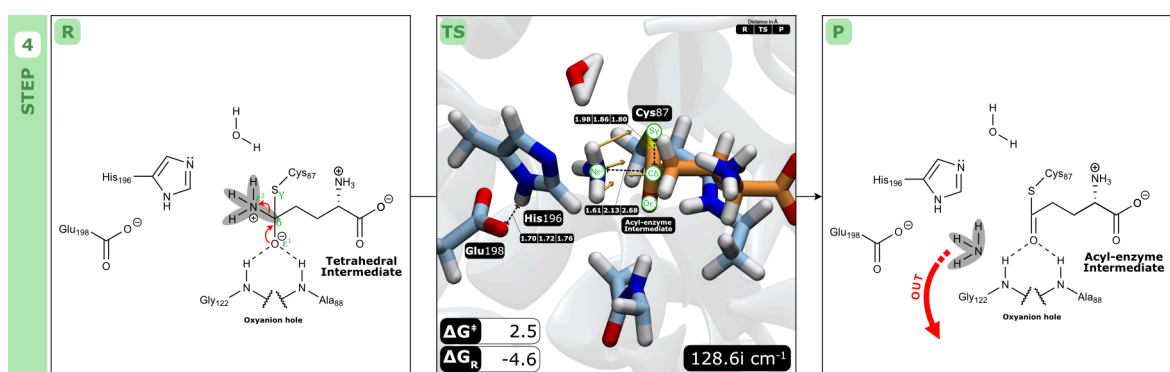


Figure 25. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 4. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. Glutamine, the substrate, is represented in orange, whereas the rest is in light blue. The yellow vectors illustrate the major vibrational modes corresponding to the imaginary frequency of the TS (shown in the bottom right corner). Relevant interatomic distances (in Å) for **R**, **TS**, **P** are shown, as well as the activation and reaction Gibbs free energies for this step.

In the reactants of this reaction, the positively charged NH₃ group of the tetrahedral intermediate remains covalently bonded to C_δ (1.61 Å), and the negatively charged O_ε¹ from the carbonyl group continues to be stabilized by the residues from the oxanion hole (distance from Gly₁₂₂: 2.08 Å and from Asp₈₈: 1.90 Å). The transition state of this reaction is characterized by an imaginary frequency at 128.6i cm⁻¹. In this stationary point, the NH₃ group is already displaced from carbon C_δ (2.13 vs 1.61 Å in the reactants), thus favouring the reaction. In the reaction product, the NH₃ group is finally released as ammonia (2.68 vs 1.61 Å in the reactants), and the carbonyl group reverts to its initial configuration (product vs reactants, distance: 1.23 vs 1.28 Å; charge of O_ε¹: -0.57 vs -0.76 a. u.; bond order: 1.75 vs 1.53). Cys₈₇ continues to be firmly attached to C_δ of the substrate (1.80 Å;

bond order: 0.99), forming the glutamyl-thioester intermediate. The typical carbon-sulfur bond distances in thioesters range from 1.74 to 1.78 Å, according to the Cambridge Data Base, which is in agreement with our results for the distance of the glutamyl-thioester intermediate (275).

The results demonstrate that this reaction is very fast, requiring a Gibbs free energy of activation of $2.5 \text{ kcal.mol}^{-1}$ and being exergonic in $4.6 \text{ kcal.mol}^{-1}$. The exergonicity of the reaction can be explained by the occurrence of a bond break ($\text{N}_\epsilon^2\text{-C}_\delta$) in the tetrahedral intermediate with the release of one ammonia molecule.

1.1.3.3. Stage 3: Hydrolysis of the glutamyl-thioester intermediate

This stage comprises the hydrolysis of the glutamyl-thioester intermediate by a nucleophilic attack to its carbon C_δ (step 5; **Figure 26**). The simulations suggest that the water molecule responsible for the hydrolysis, which is widely available in the active site due to easy access to the solvent, is activated by His_{196} , which acts as a base, abstracting one of its protons. It was postulated that the highly conserved crystallographic water molecule present in the active site is the one involved in the hydrolysis (34). However, its position in our model is not concordant with this role.

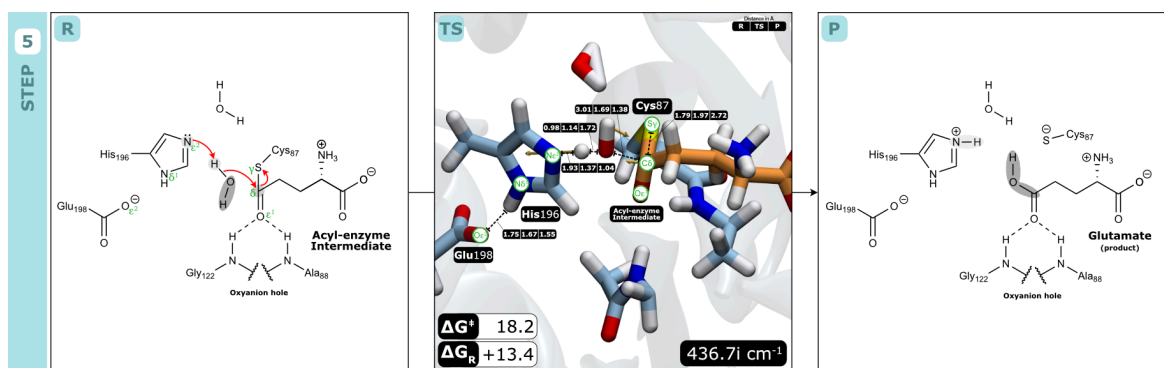


Figure 26. Structures of reactant (R), transition state (TS), and product (P) of step 5. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. Glutamine, the substrate, is represented in orange, whereas the rest is in light blue. The yellow vectors illustrate the major vibrational modes corresponding to the imaginary frequency of the TS (shown in the bottom right corner). Relevant interatomic distances (in Å) for R, TS, P are shown, as well as the activation and reaction Gibbs free energies for this step.

In the reactants, the proton of the water molecule that points towards the nitrogen N_{ϵ}^2 of His₁₉₆ is at 1.93 Å, while its oxygen atom is at 3.01 Å from the carbon C_{δ} of the glutamyl-thioester intermediate.

The transition state determined for this reaction is characterized by an imaginary frequency at 436.7i cm^{-1} . It is clear that several simultaneous reactions occur in this step. While the water molecule gets closer to carbon C_{δ} of the intermediate [distance (bond order): 1.69 Å (0.52) vs 3.01 Å (0.03) in the reactants], Cys₈₇ starts to dissociate from the same carbon [1.97 Å (0.67) vs 1.79 Å (1.01) in the reactants], and the bond length of the carbonyl group increases [1.26 Å (1.61) vs 1.22 Å (1.77) in the reactants].

In the reaction product, His₁₉₆ becomes cationic once again (0.75 a. u. vs 0.02 a. u. in the reactants), and the carbon C_{δ} of the now-formed glutamate becomes covalently bonded to the hydroxyl group derived from the water molecule (1.38 Å vs 3.01 Å in the reactants).

The calculations indicate that this reaction requires a Gibbs free energy of activation of 18.2 kcal.mol^{-1} and is endergonic in 13.4 kcal.mol^{-1} . The high reaction energy can be justified by the formation of two point charges in the residues Cys₈₇ and His₁₉₆. The results show that this reaction is accomplished in a single step and, contrarily to what is proposed in the literature for other amidases (e. g. Nitrilase 2) (205), there is no formation of a second tetrahedral intermediate upon reaction with the water molecule.

1.1.3.4. Stage 4: Enzymatic turnover

Once the product of the reaction is released from the active site, His₁₉₆ becomes positively charged and Cys₈₇ anionic. From this point, it is hypothesized that two situations can occur: i) if substrate concentration is high enough, it can bind to the active site and react directly with the anionic Cys₈₇, continuing the catalytic cycle. Otherwise, ii) the enzyme can return to its steady-state in which Cys₈₆ becomes protonated and His₁₉₆ neutral (step 6; **Figure 27**).

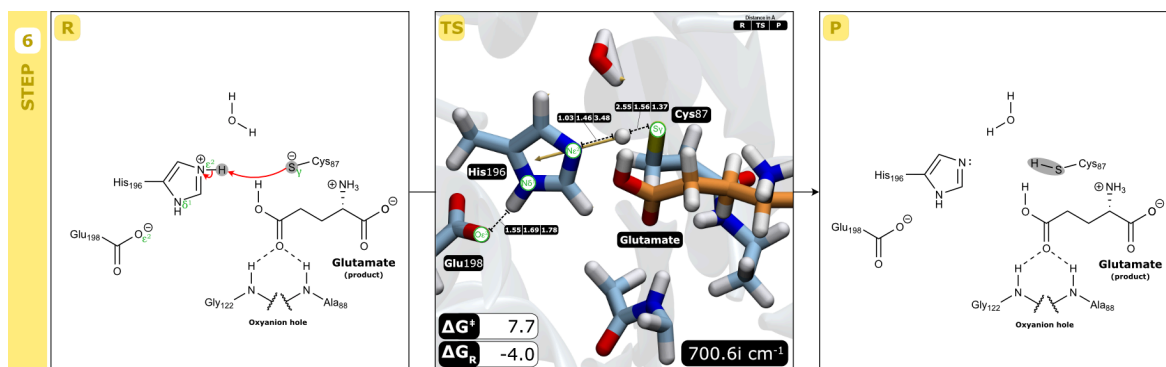


Figure 27. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 6. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. Glutamine, the substrate, is represented in orange, whereas the rest is in light blue. The yellow vectors illustrate the major vibrational modes corresponding to the imaginary frequency of the TS (shown in the bottom right corner). Relevant interatomic distances (in Å) for **R**, **TS**, **P** are shown, as well as the activation and reaction Gibbs free energies for this step.

The computational results have shown that such reaction requires a Gibbs free energy of activation of $7.7 \text{ kcal.mol}^{-1}$ and is exergonic in $4.0 \text{ kcal.mol}^{-1}$. The transition state obtained for this reaction is characterized by an imaginary frequency at $700.6i \text{ cm}^{-1}$.

Considering the energetic cost of this reaction, this is indicative that it only occurs when no substrate is present near the active site. The deactivation of Cys₈₇ described in stage 4 is presumed to occur only when the substrate concentration near the active site is low. Otherwise, if enough substrate is available near the active site, since that Cys₈₇ is already in thiolate form, the reaction can swiftly proceed. This way, the reaction described in this stage works like a protective mechanism that protects the reactive thiolate of Cys₈₇ from other molecules that could potentially harm or damage the enzyme's active site, culminating in its inhibition. This type of activation mechanism was already observed in other enzymes (205,306).

1.1.3.5. Energy Profile

The overall Gibbs free energy profile determined for the reaction catalyzed by Pdx2, the glutaminase domain of Pyridoxal 5'-phosphate synthase, is shown in **Figure 28**, summing up the results obtained. Four distinct stages can be identified: activation, deamination, hydrolysis, and enzymatic turnover. In the first stage occurs the activation of Cys₈₇ by the removal of a proton from His₁₉₆, forming a thiolate which is a better nucleophile for the

following step. In the second stage, three steps occur, starting with the nucleophilic attack of the thiolate form of Cys₈₇ to the carbon C_δ of the amide group of glutamine, which is simultaneous with the formation of the tetrahedral intermediate, that becomes covalently attached to the enzyme. The second step of stage two (step 3) is almost spontaneous and consists in the protonation of the NH₃ group of the tetrahedral intermediate by His₁₉₆. In the next step (step 4), this group is released as ammonia to the active site, resulting in the covalently attached glutamyl-thioester intermediate. In the third stage, the hydrolysis of the glutamyl-thioester intermediate occurs in a single step. This reaction is the rate-limiting step of the entire catalytic mechanism and requires a Gibbs free energy of activation of 18.2 kcal.mol⁻¹.

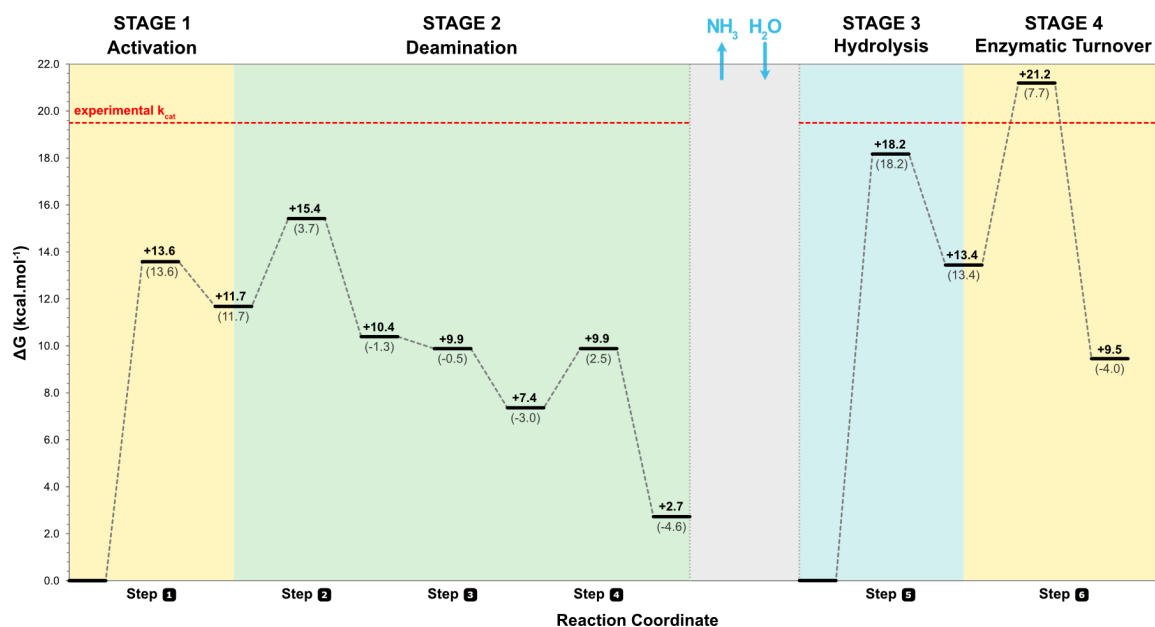


Figure 28. Energy profile determined for the catalysis of the conversion of glutamine into glutamate by Pdx2, with the release of ammonia. The values in parentheses correspond to the corresponding activation or reaction Gibbs free energy of each step. All energy values were computed with DLPNO-CCSD(T)/CBS//B3LYP/6-31G(d,p):ff14SB.

Once hydrolysis is complete, the mechanism can follow either of two directions. If the substrate concentration near the active site is low, the enzyme goes back to its initial state, where Cys₈₇ is protonated and His₁₉₆ neutral (stage 4). Otherwise, a new molecule of substrate can enter the active site and react with the anionic thiolate of Cys₈₇ and start a new catalytic cycle.

The available kinetic data for Pdx2 show turnover numbers that range between 0.08 and 0.13 s⁻¹ (28,33,40,42), which, considering the transition state theory, corresponds to an

energy between 19.1 to 19.5 kcal.mol⁻¹. Based on the full energetic profile of the catalytic mechanism, the rate-limiting step is the combination of steps 5 and 6, i. e., the hydrolysis and the turnover steps, which accounts for 21.2 kcal.mol⁻¹. However, based on what was discussed above, when substrate is in excess, the rate-limiting step should solely depend on the hydrolysis stage, which requires 18.2 kcal.mol⁻¹, closely resembling the experimental k_{cat} value.

1.1.4. Conclusion

In this work, we presented, with atomistic detail, the catalytic mechanism of Pdx2, as well as its associated Gibbs free energy profile studied using the computational hybrid QM/MM methodology, ONIOM. Pdx2, the glutaminase subunit of PLP synthase, catalyzes the formation of glutamate from glutamine. During this process, Pdx2 releases an ammonia molecule that is then used by the adjacent Pdx1 subunit as the source of the ring nitrogen of PLP, being essential for its synthesis.

The complete catalytic mechanism of Pdx2, summarized in **Figure 29**, occurs in six sequential steps that can be divided into four stages: (i) activation of Cys₈₇, (ii) deamination of glutamine with the formation of the glutamyl-thioester intermediate, (iii) hydrolysis of the formed intermediate, and (iv) enzymatic turnover.

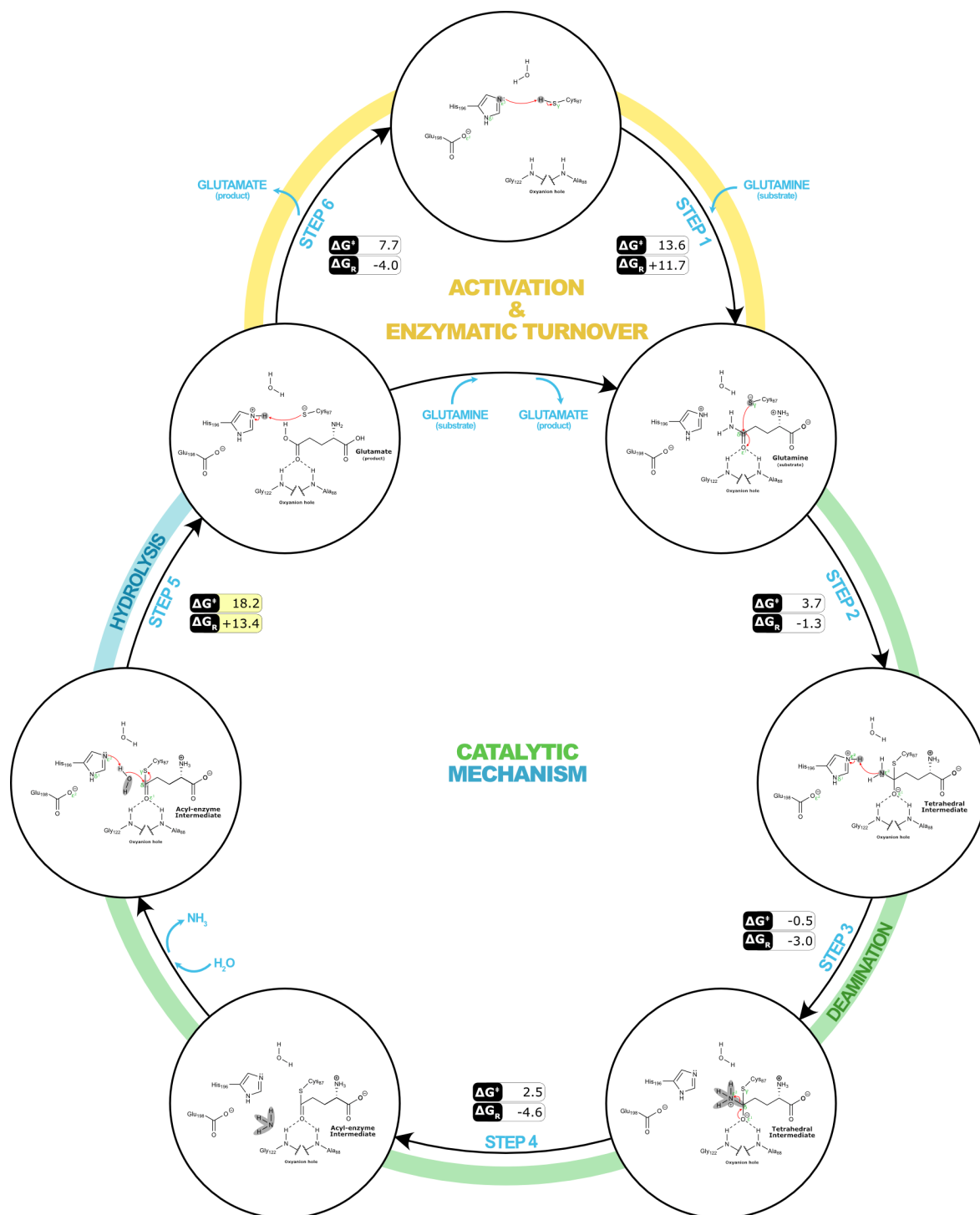


Figure 29. Overview of the catalytic mechanism of Pdx2, with the indication of the distinct phases: **Activation & Enzymatic Turnover**, **Deamination**, and **Hydrolysis**. The Gibbs free energies of activation ($\Delta G^\ddagger$) and reaction (ΔG_R) are expressed in kcal.mol⁻¹.

In the first step, Cys₈₇ is activated by the abstraction of one proton by His₁₉₆, forming a thiolate, a better nucleophile than the thiol. Contrarily to what is hypothesized in the literature (34), the conserved crystallographic water molecule present in the active site

does not have a catalytic role in the first steps of catalysis, potentially suggesting that its role is only to correctly orientate both Cys₈₇ and the imidazole ring of His₁₉₆ for the reaction to happen. Moreover, the role of the Glu₁₉₈, part of the catalytic triad, was discussed. In contrast to what is verified in other enzymes employing a classic catalytic triad (Ser-His-Glu), Glu₁₉₈ does not have an active role in the catalysis since it does not receive a proton from His₁₉₆. The role of Glu₁₉₈ appears to be linked with the correct positioning and stabilization of the imidazole ring of His throughout the reaction, but it is not directly involved in the catalytic process.

At the beginning of the second stage, Cys₈₇ undergoes a nucleophilic attack to the carbon C₅ of the substrate, glutamine, resulting in the formation of the tetrahedral intermediate, which is covalently bonded to the enzyme. Then, one proton is transferred from His₁₉₆ to the amino group of this intermediate, forming a transient -NH₃⁺ group that is quickly released in the form of ammonia at the end of stage 2. This results in the formation of the glutamyl-thioester intermediate, still covalently bonded to the enzyme.

In the third stage, the glutamyl-thioester intermediate is hydrolyzed in a single step, freeing Cys₈₇ and forming the final product, glutamate. This is another difference in the catalytic mechanism of Pdx2 in comparison with other amidases, such as the Malonamidase E2 (307), where the hydrolysis of the acyl-enzyme intermediate adduct occurs in two steps and involves the formation of a second tetrahedral intermediate.

Lastly, at the end of the catalytic process, two hypotheses can occur: (a) if enough substrate is readily available, a new substrate molecule can enter the active site, and the catalytic cycle proceeds with the already deprotonated Cys₈₇ and protonated His₁₉₆, needed for the catalysis to occur; or (b) if the substrate concentration near the active site is low, the enzyme can return to its initial state, with Cys₈₇ protonated and His₁₉₆ neutral (step 6). The relatively high energy barrier of the last step (7.7 kcal.mol⁻¹) indicates that it would be inefficient for the enzyme to reprotonate Cys₈₇ only to deprotonate it right after. Therefore, it is suggested that this reaction works as a protective mechanism against other reactive molecules that could interact with the activated Cys₈₇ rendering the enzyme inactive. This improves the enzyme's catalytic efficiency since it does not have the energy penalty of reprotonation/deprotonation of Cys₈₇.

This activation/deactivation mechanism is only possible because Cys₈₇ can be found in the anionic form at physiological pH (pK_a≈8). On the other hand, amidases that have a Ser residue (pK_a≈13) instead of a Cys in the catalytic triad, such mechanism would not be possible. The rate-limiting step of the full catalytic process, when the substrate is in excess (**Figure 28**), is the hydrolysis of the glutamyl-thioester intermediate (step 5), which

requires an activation energy of $18.2 \text{ kcal.mol}^{-1}$, agreeing with the available kinetic data that predicted an activation barrier in the range of $19.1\text{--}19.5 \text{ kcal.mol}^{-1}$ (28,33,40,42).

In this study, it was also clear that the correct orientation and positioning of the residues of the active site of Pdx2 is very important to trigger the catalytic process once the substrate binds in the active site. This is observed in all the transition states of the calculated reactions where the Bürgi-Dunitz (BD) and the Flippin-Lodge (FL) angles are very close to the optimal values (BD: 107° and FL: 0°) that favour nucleophilic additions and elimination reactions. This is particularly evident in the nucleophilic attack of Cys₈₇ to the substrate or the hydrolysis of the glutamyl-thioester intermediate enzymatic adduct.

The results obtained in this study provide a detailed atomistic description of the catalytic mechanism of Pdx2 with all the chemical steps required for the transformation of glutamine into glutamate. Furthermore, taking into account that most of the inhibitors targeting this enzyme are irreversible inhibitors that become irreversibly attached to the enzyme through Cys₈₇, this article provides for the first time details that can be used to develop new compounds that react faster with the enzyme and form a tetrahedral reaction intermediate that could irreversibly inactive the enzyme, hindering the hydrolysis process.

2. PLP-dependent enzymes

This page was intentionally left blank.

2.1. Catalytic mechanism of XometC – a β -elimination reaction

XometC is a pyridoxal 5'-phosphate-dependent enzyme from the pathogenic bacteria *Xanthomonas oryzae pv. oryzae*, the causing agent of the bacterial blight of rice, one of the most destructive diseases that affect rice production worldwide. For this reason, XometC has been regarded as a potential drug target for the development of new bactericides against this disease. XometC was shown to be a promiscuous enzyme presenting both cystathionine γ -lyase and L-serine ammonia-lyase activity. In this work, the catalytic mechanism of its L-serine ammonia-lyase activity, a β -elimination reaction, was studied using computational quantum mechanics/molecular mechanics methodologies. XometC is responsible for the metabolism of the amino acid L-serine into 2-aminoacrylate, which in turn is converted to pyruvate in the cytosol. The results have shown that its catalytic mechanism takes place in 14 sequential steps that go from the formation of the external aldimine with L-serine, the β -elimination, the re-formation of the internal aldimine, and the enzymatic turnover. Furthermore, the determined Gibbs free energy profile shows that the elimination of the hydroxyl group of L-serine is the rate-limiting step, with an energy barrier of 19.1 kcal.mol⁻¹, which closely agrees with the experimental kinetic data ($k_{cat} = 8.5 \pm 0.4 \text{ min}^{-1}$, corresponding to $\sim 18.9 \text{ kcal.mol}^{-1}$). Together, these results provide insight into the catalytic mechanism of XometC, which can aid in developing new and selective drugs targeting this enzyme with minor toxicity for humans.

André F. Pina performed all the calculations and, considering all co-authors' insights, wrote the article, which is currently under submission.

2.1.1. Introduction

XometC is a PLP-dependent enzyme from *Xanthomonas oryzae* pv. *oryzae* (Xoo), a pathogenic bacterium responsible for rice bacterial blight (BB). BB is a devastating disease that affects most rice-growing countries, being capable of destroying up to 80 per cent of a crop resulting in grave rice production constraints with a high impact on food availability (308,309).

XometC is essential in the amino acid metabolism of Xoo, accepting substrates like L-cystathionine, L-O-succinylhomoserine, L-homoserine, and L-serine. The inhibition of this enzyme can impair the activity of Xoo, which makes it a potential drug target in the development of novel bactericides to fight BB (308).

XometC was initially crystallized by Ngo et al. (71,72) and characterized as a homotetramer with four active sites composed of residues from every two adjacent protomers. This enzyme is described as having two different catalytic activities: cystathionine γ -lyase-like activity (EC 4.4.1.1) and L-serine ammonia-lyase activity (EC 4.3.1.17; originally denominated L-serine dehydratase) (72). Nevertheless, compared with other L-serine ammonia-lyases, its active site presents significant differences, as can be noticed in the crystallographic structures from *Dictyostelium discoideum* (PDB 6LUT) (310), *Rhizomucor miehei* CAU432 (PDB 5C3U) (311), *Rattus norvegicus* (PDB 1PWH) (312), and *Homo sapiens* (PDB 4H27 and 1P5J) (313,314). The main difference between its active site and the others is in the residue that interacts with the nitrogen N1 of the pyridine ring of the PLP cofactor. While the active site of other L-serine ammonia-lyases contains either a serine (PDB 6LUT), a threonine (PDB 5C3U), or a cysteine (PDB 1PWH, 4H27, and 1P5J) residues, XometC has an aspartate. Although in both cases, the nitrogen N1 establishes one hydrogen bond with these residues, they drastically alter the properties of the PLP cofactor. In the case of the active site of XometC, the nitrogen N1 is protonated, and the aspartate residue is proposed to help with the stabilization of negatively charged PLP reaction intermediates (315). On the other hand, in the L-serine ammonia-lyase active site, this process is not possible.

Another key difference between both active sites is at the re-face of PLP. In the case of the XometC active site, there is a tyrosine residue, Tyr₁₁₂, that donates a proton to the substrate and promotes the β -elimination reaction of the hydroxyl group. On the other hand, in the L-serine ammonia-lyase active site, this position is occupied either by a glycine (PDB 6LUT), a serine (PDB 5C3U), or an alanine (PDB 1PWH, 4H27, and 1P5J). This indicates that the β -elimination catalyzed by both active sites undergoes different catalytic pathways.

Taking this into account, the catalytic mechanism of the L-serine ammonia-lyase activity of the XometC enzyme (**Figure 30**) was studied with atomistic detail using computational methods, namely quantum mechanics/molecular mechanics (QM/MM). The studied reaction is divided into three main stages and involves the activation of the enzyme, the transamination, and the beta-elimination reactions. The activation stage involves the formation of a covalent bond between the PLP cofactor and an active site lysine, generating what is called the internal aldimine. The transamination stage involves the transformation of the internal aldimine into the external aldimine EAS, in which the L-serine substrate becomes covalently bonded to the PLP cofactor. The next stage involves the beta-elimination process, which results in the external aldimine EAA, which contains the reaction's product (2-aminoacrylate) connected to the PLP cofactor. From this point, the reverse reaction of the transamination occurs through the conversion of the external aldimine EAA to the internal aldimine. The enzyme will remain in this state if new substrates are available for a new catalytic cycle. Otherwise, it will revert to its inactive state.

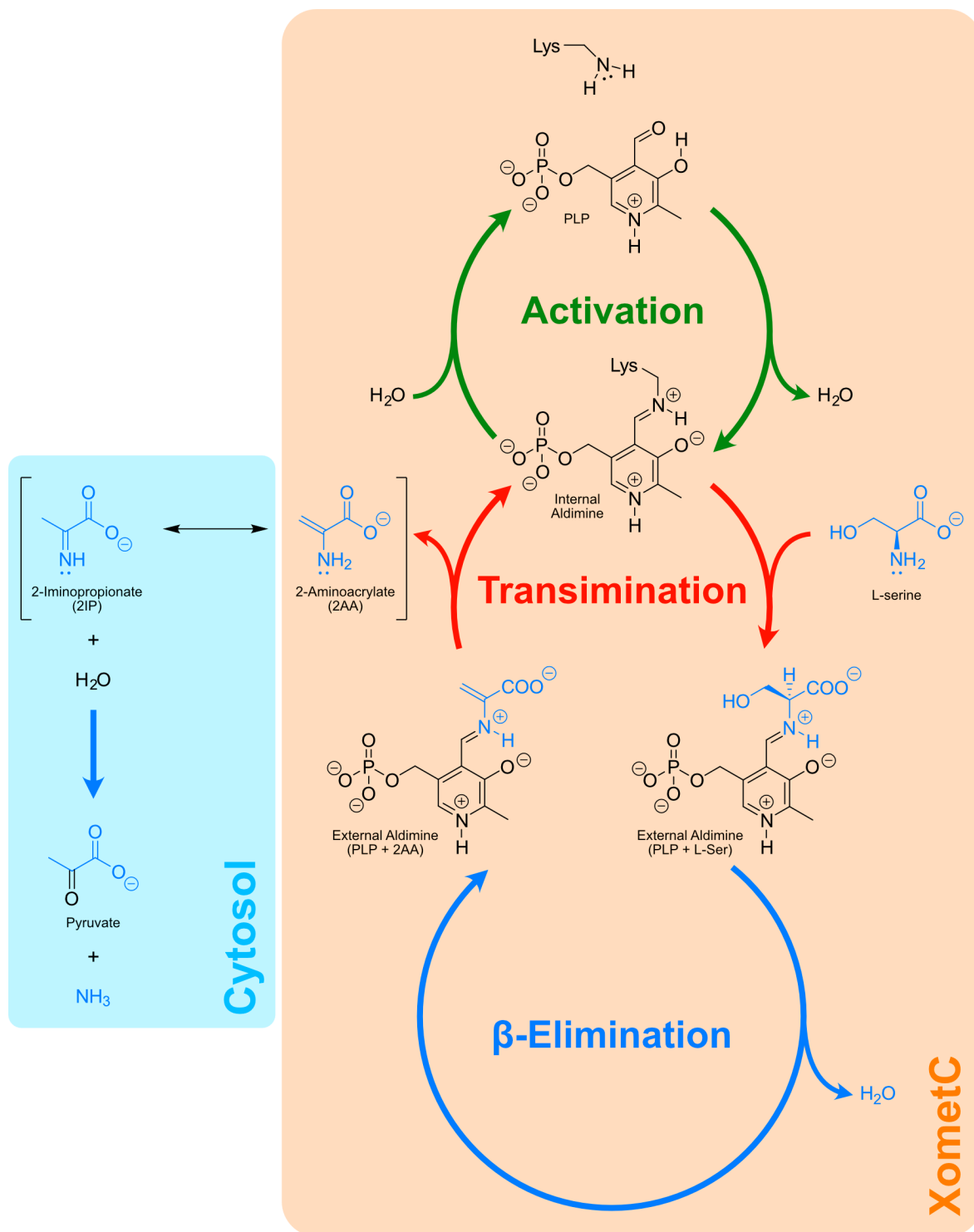


Figure 30. Schematic representation of the catalytic mechanism of XometC, a promiscuous PLP-dependent enzyme able to catalyze the β -elimination of the hydroxyl group of L-serine with the formation of 2-aminoacrylate, later converted into pyruvate in the cytosol. The formation of the internal and external aldimines is also depicted.

2.1.2. Methodology

2.1.2.1. Model preparation

The model used in this study was built based on the X-ray crystallographic structure of the homotetramer cystathionine γ -lyase-like protein (hereby referenced as XometC) from *Xanthomonas oryzae pv. oryzae* KACC 10331 (PDB: 4IY7; resolution: 1.70 Å) (72). The crystallization artefacts, glycerol, pyruvic acid, and sulfate ions were removed. The external aldimines containing the pyridoxal 5'-phosphate (PLP) bonded to the substrate L-serine (EAS) present in three of the four active sites were kept, being the starting point of our study. The external aldimine of PLP with the product 2-aminoacrylate (EAA), present in the other active site, was also kept and was used to compare with the results of the QM/MM studies with the same reaction intermediate. There is also a zwitterionic L-serine amino acid in a binding site close to the active site, which was also kept.

The protonation state of all residues at the physiological pH of 7.4 was estimated in the ProteinPrepare web server (316) with EAS or EAA present in XometC's active sites.

An additional model was also built to study the non-enzymatic conversion of the product, 2-aminoacrylate (2AA), into pyruvate. This model contained one 2AA molecule and two water molecules. In order to mimic the solvent effect, the implicit solvent model SMD (317) was used as implemented in the ORCA software, version 5.0.3 (295,296,318–320). Water ($\epsilon_r = 78.3550$, $n = 1.3328$) was used as a solvent.

2.1.2.2. Structure minimization

In order to relax possible strains that might have resulted from the model preparation, a series of energy minimization steps were carried out with the AmberTools software package, version 19 (321). The ff14SB force field (129) was used for the protein, while the second generation of the General Amber Force Field (GAFF2) (132,133) was used for the ligands EAS and EAA. To obtain the partial atomic charges for these ligands, the Restrained ElectroStatic Potential (RESP) charge derivation protocol was used (322–324). These charges were calculated from the respective optimized structure at the Hartree-Fock level of theory with the 6-31G(d) basis set (i.e., HF/6-31G(d)). The geometry optimization was performed using Gaussian09 (284). The zwitterionic L-serine amino acid parameters were obtained from the parameter set for zwitterionic amino acid residues from Horn (325).

The system was solvated in a box of TIP3P (286) water molecules with a prismatic shape, whose sides were at least 12 Å from any atom of the protein. Then, to have a physiological concentration of NaCl (i.e., 150 mM), Na⁺ and Cl⁻ ions were added to the box in random positions. The number of Na⁺ and Cl⁻ ions added was determined using the SPLIT method (326). All residues were protonated by the LeaP package of the AmberTools software according to titrated pK_a obtained as mentioned in **section 2.1.2.1**.

The minimization procedure consisted of four stages. In each stage, the first half of the total steps use the steepest descent method, then change to the conjugate gradient method. Firstly, only the water molecules were minimized (5000 steps), with cartesian positional restraints with a weight of 50 kcal.mol⁻¹.Å² applied on all other atoms. Secondly, all hydrogen atoms were minimized (5000 steps), with the same restraint weight applied to all non-hydrogen atoms. Thirdly, all non-backbone atoms in the system were minimized (5000 steps), and the same restraint weight was applied to all protein backbone atoms. Lastly, all system atoms were minimized with no restraints (100 000 steps).

2.1.2.3. QM/MM

The study of the catalytic mechanism of XometC was carried out through quantum mechanics/molecular mechanics (QM/MM) calculations using the additive scheme with electrostatic embedding. The ORCA software package, version 5.0.3 (295,296,318–320), was used to perform these calculations.

The QM/MM model used in this study was built from the system structure obtained after the fourth step of minimization with the aid of the molUP (257) plugin for the Visual Molecular Dynamics (VMD) software (287), obtained through VMD Store (288).

The model (**Figure 31**) consisted of the entire XometC homotetramer, with a water shell of 6 Å that was kept around it. In total, the model system contains 40 488 atoms. This model was then divided into two regions treated at different theoretical levels: the quantum mechanics (QM) region and the molecular mechanics (MM) region. The QM region contains 192 atoms belonging to chain C unless stated otherwise. This region included the PLP bonded to L-serine (EAS), the zwitterionic L-serine, a water molecule, Tyr₁₁₂, Gly₁₁₃, Gly₁₁₄, Thr₁₁₅, and the side chains of Asn₁₆₀, Asp₁₈₅, Thr₁₈₇, Lys₂₁₀, Glu₃₃₈, Ser₃₃₉, and Arg₃₇₄. In addition, the side chains of the residues stabilizing the phosphate group of PLP, Tyr₅₈ (chain D), Arg₆₀ (up to its CD carbon; chain D), Thr₂₀₉, and Ser₂₀₇, were also included. The link atom approach was used to complete the valences of the bonds between the two regions by adding a hydrogen atom along the bond vector severed by

the QM/MM interface. Due to the great size of this system, all residues that were farther than 2 Å from the QM region were frozen throughout the calculations to reduce the computational cost.

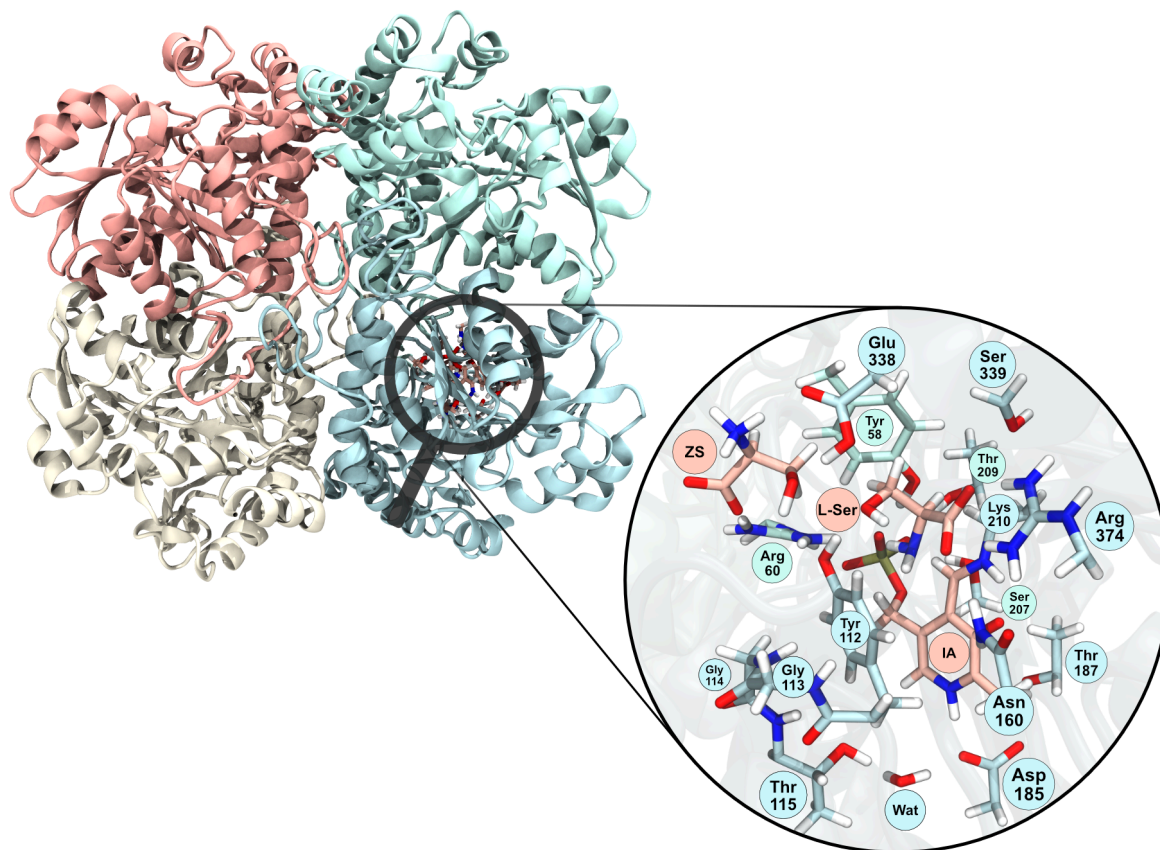


Figure 31. (Left) Structure of the XometC homotetramer, with each subunit coloured differently. The residues belonging to the QM region of the bottom right subunit are represented in Licorice. (Right) PLP, the substrate, and the zwitterionic L-serine are coloured pink, while the active site residues are coloured blue. The residues Tyr₅₈, Arg₆₀, and Thr₂₀₉ belong to the adjacent subunit. Each molecule is represented in Licorice and is identified by a circular label.

The QM/MM calculations employed in this study followed a custom workflow, allowing us to explore the reactive conformational space of the enzyme's active site more efficiently in the same timeframe. To do so, in each step, we started by performing an initial optimization with GFN2-xTB (255), a semiempirical tight-binding method primarily designed for the fast calculation of biochemical systems with a few thousand atoms while yielding reasonable starting structures. If the system was far from convergence, we first performed a GFN2-xTB optimization using the very fast and efficient L-BFGS (327) optimizer (200-500 iterations) and then changed to the default Rational Function Optimization (RFO) method (328). Subsequently, still using GFN2-xTB, we performed a relaxed surface scan through the coordinate of interest until after finding the maximum of

the curve, a good guess for the transition state. From this curve, we selected the two points closer to each side of the maximum and optimized them freely to obtain the reactant and product of the reaction. These minima are then reoptimized with the composite electronic-structure method r^2 SCAN-3c (247), a recently developed meta-generalized-gradient approximation (mGGA) method, together with a modified version of the def2-TZVP (199) basis set denominated mTZVPP (247). The r^2 SCAN-3c is a composite electronic-structure method that has shown comparable results with standard hybrid-DFT/QZ methods at a computational cost 100-1000 times lower (228,247,329).

Then, to find the transition state structure, we used an implementation of the Nudged Elastic Band (NEB) method (330–332) in ORCA called NEB-TS (333). This method is the most efficient in the search of a saddle point, combining the NEB with a subsequent eigenvector-following calculation. The r^2 SCAN-3c was used to describe the QM region.

Vibrational frequency calculations were performed to validate the transition state by the existence of only one imaginary frequency associated with the reaction coordinate. Since these calculations quickly become very computationally demanding for larger systems, the semiempirical GFN2-xTB functional was used (228,255,334). Vibrational frequency calculations were also performed for all structures (reactants, transition states, and products), and only one imaginary frequency was found for each transition state. The zero-point energy and thermal, enthalpy, and entropic corrections were estimated at 303.15 K and 1.0 atm, the same conditions used in the experimental k_{cat} determination.

Finally, single-point (SP) energy calculations were performed on the previously r^2 SCAN-3c optimized structures (i.e., reactants, transition states, and products) to obtain the final energy values. The DLPNO-CCSD(T)/CBS:ff14SB level of theory was employed. The Domain-based Local Pair Natural Orbital Coupled-Cluster method with Single, Double, and perturbative Triple excitations (DLPNO-CCSD(T)) was used since it combines the high accuracy of CCSD(T), the gold-standard method, with a computational cost similar to DFT (219–222). Automatic extrapolation to the basis set limit was performed in ORCA using the keyword Extrapolate(2/3,cc). This indicates that a two-point extrapolation was employed using cc-pVDZ and cc-pVTZ basis sets. In addition, the cc-pVTZ/C auxiliary basis set (298), required for correlation fitting, was used. The RI-JK approximation and the required cc-pVTZ/JK auxiliary basis set were employed to accelerate the SCF calculation.

The final energy values are a sum of the electronic energies of the QM region from the DLPNO-CCSD(T)/CBS calculation, the MM contributions previously calculated with the ff14SB force field during geometry optimization, and the thermochemical corrections obtained in the vibrational frequency calculations.

The net atomic charges presented throughout this work were obtained through the software Multiwfn, version 3.8 (335), by employing the atomic dipole corrected Hirshfeld population method (336).

2.1.2.4. Simulating UV-Vis spectra

UV-Vis spectra were simulated for the following structures: internal aldimine, external aldimine with the substrate L-serine, the quinonoid intermediate, and external aldimine with the product 2-aminoacrylate. After optimizing the ground state of these structures using the entire model of the enzyme with QM/MM methods, we performed single point excited state energy calculations on the QM layer using the time-dependent density functional theory (TD-DFT) approach in ORCA, version 5.0.3. The PBE0 (337,338) functional was used as it has a good balance between predictive performance and computational cost, therefore being a commonly used functional in the calculation of excited states with TD-DFT. The basis set used was def2-SV(P) (199), as it is not very expensive and provides enough accuracy for this purpose. The Tamm-Dancoff approximation (TDA) was turned off since it reduces the accuracy of the oscillator and rotatory strengths, not being suitable for the simulation of UV-Vis spectra.

2.1.3. Results and Discussion

The enzyme XometC is only active when PLP is covalently bonded to Lys₂₁₀ by the carbon C4A (i.e., in its internal aldimine (IA) form). For this reason, the catalytic mechanism was studied from this step onward.

Therefore, the computational calculations regarding the catalytic mechanism of XometC were divided into four stages: i) formation of external aldimine with L-serine; ii) β -elimination; iii) re-formation of the internal aldimine; iv) reprotonation of Tyr₁₁₂ and subsequent enzymatic turnover. All these stages require fourteen sequential steps, which will be described and discussed in detail in the following sections.

2.1.3.1. Stage 1: Formation of external aldimine with L-serine

The conversion of L-serine into 2-aminoacrylate catalyzed by XometC starts with the formation of an external aldimine with L-serine (EAS), in which L-serine (L-Ser) becomes covalently bonded to PLP. This is accomplished in four sequential steps.

In the first step, the nitrogen N1 of the amino group of L-Ser performs a nucleophilic attack to carbon C4A of the IA. This leads to the formation of a gem-diamine reaction intermediate (GS1) in which both L-Ser and Lys₂₁₀ are covalently bonded to PLP. As such, the bond between the N_ζ of Lys₂₁₀ and the C4A of PLP loses the double-bond character (R: 1.30 Å to P: 1.40 Å), as depicted in **Figure 32**. This reaction requires a Gibbs free energy of activation of 3.2 kcal.mol⁻¹ and is exergonic in 1.2 kcal.mol⁻¹. The TS structure is characterized by a small imaginary frequency at 87.6i cm⁻¹ associated with the vibration of the atoms involved in the bond formation.

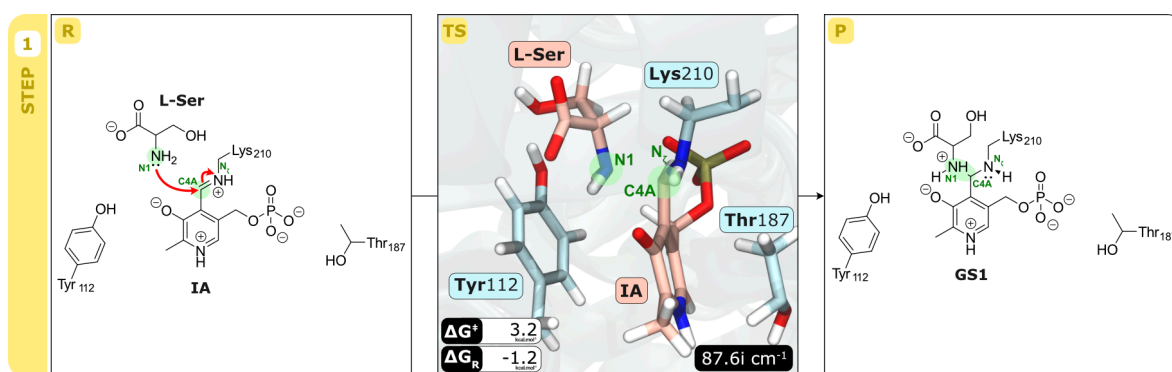


Figure 32. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 1. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. PLP and L-Ser are represented in pink, while the rest are in light blue. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown.

The second step involves a small rearrangement in the network of H-bonds in the active site (**Figure 33** – Step 2). In the reactant, the nitrogen N_ζ from Lys₂₁₀ establishes one H-bond with the oxygen O3 of PLP (2.10 Å). Its position then shifts to establish a new H-bond with the oxygen OH of Ser₂₀₇ (2.06 Å). This allows the nitrogen N1 of L-Ser to establish a stronger H-bond with the oxygen O3 of PLP (R: 2.01 Å vs P: 1.68 Å), becoming better oriented for the next step to occur. This step has a Gibbs free energy of activation of 3.3 kcal.mol⁻¹ and is slightly exergonic in 0.7 kcal.mol⁻¹. The TS structure is characterized by one imaginary frequency at 430.8i cm⁻¹.

In step 3 (**Figure 33**), a proton transfer from the nitrogen N1 of L-Ser to the oxygen O3 of PLP occurs. This step is virtually spontaneous, requiring a Gibbs free energy of activation of $-0.4 \text{ kcal.mol}^{-1}$, and the reaction is exergonic in $2.8 \text{ kcal.mol}^{-1}$. The TS of this reaction is characterized by one imaginary frequency at $644.8i \text{ cm}^{-1}$.

The last step of this stage, step 4, involves the cleavage of the bond between the carbon C4A of PLP and nitrogen N_{ζ} of Lys₂₁₀. However, this reaction requires a previous proton transfer between atoms N1 of L-Ser and N_{ζ} of Lys₂₁₀ through the oxygen O3 of PLP (**Figure S01**). After numerous tries, we could not characterize its TS since its energy barrier must be considerably low. Nevertheless, the product of this reaction was characterized and is $5.8 \text{ kcal.mol}^{-1}$ more stable than the product from step 3.

At the end of step 4, a double bond between the nitrogen N1 of L-Ser and the carbon C4A of PLP is formed (R: $1.41 \text{ \AA}$ vs P: $1.31 \text{ \AA}$) (**Figure 33 – Step 4**), from which results the formation of the external aldimine EAS. This step required a Gibbs free energy of activation of $4.2 \text{ kcal.mol}^{-1}$ and is exergonic in $3.9 \text{ kcal.mol}^{-1}$. Its TS is characterized by one imaginary frequency at $105.2i \text{ cm}^{-1}$.

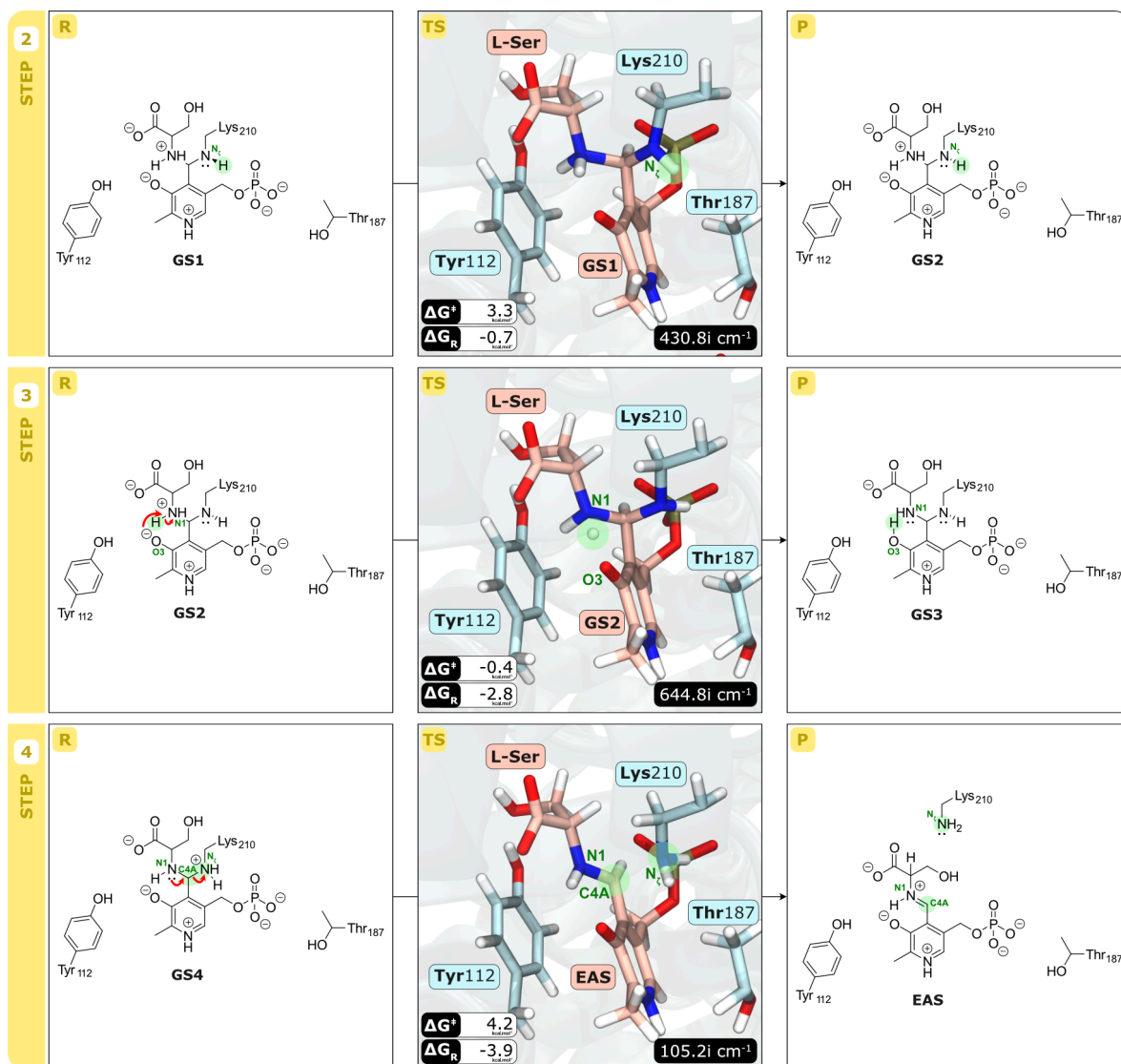


Figure 33. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of steps 2-4. The structures of the reactants and products are represented in a wedge-dash manner. The structures of the TS are represented three-dimensionally, using the Licorice drawing method of VMD. PLP and L-Ser are represented in pink, while the rest are in light blue. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown.

2.1.3.2. Stage 2: β -Elimination

After the formation of the EAS, the enzyme is ready to catalyze the β -elimination of the hydroxyl group of L-Ser to form 2-aminoacrylate (2AA). This stage occurs in two steps: i) abstraction of the proton from C $_{\alpha}$ of L-Ser (Step 5) and ii) elimination of the hydroxyl group (Step 6).

Step 5 is characterized by a proton transfer between carbon C $_{\alpha}$ of L-Ser and the nitrogen N1 of the neutral Lys₂₁₀ (**Figure 34** – Step 5). This leads to the formation of a quinonoid intermediate (QUI) with the PLP cofactor functioning as an electron sink, stabilizing the negative charge generated in its pyridine ring (charge of PLP's ring: R: 0.14 a. u. vs P: -0.11 a. u.). This reaction requires a Gibbs free energy of activation of 8.9 kcal.mol⁻¹ and is endergonic in 2.1 kcal.mol⁻¹. The TS structure is characterized by one imaginary frequency at 1100.6i cm⁻¹ and is associated with the vibration of the atoms involved in this reaction. In the TS structure, the bond that is cleaved (i.e., the bond between the proton and the carbon C $_{\alpha}$ of L-ser) is almost perpendicular to the PLP's pyridine ring (H-C $_{\alpha}$ -C7-C2 angle: 108.4°). This condition is essential for the PLP cofactor chemistry to kinetically favour the reaction (69,70).

The second step of stage II (Step 6) is characterized by a proton transfer from the oxygen O1 of Tyr₁₁₂ to oxygen O $_{\gamma}$ of L-Ser. This reaction is concomitant with the cleavage of the bond between the oxygen O $_{\gamma}$ and the carbon C $_{\beta}$ (i.e., the β -elimination reaction) (**Figure 34** – Step 6). This leads to a charge delocalization from the pyridine ring of PLP (charge of PLP's ring: R: -0.10 a. u. vs P: 0.06 a. u.) to form a double bond between the carbons C $_{\alpha}$ and C $_{\beta}$ of the now formed external aldimine with 2-aminoacrylate (EAA) (R: 1.49 Å vs P: 1.34 Å). At the end of this reaction, one water molecule is formed, Tyr₁₁₂ becomes deprotonated, and the product 2-aminoacrylate is obtained, still covalently bonded to PLP. This reaction has the highest energy barrier, requiring a Gibbs free energy of activation of 17.0 kcal.mol⁻¹ and being endergonic in 5.5 kcal.mol⁻¹. As seen in the previous step, in the TS structure (621.6i cm⁻¹), the cleaved bond is also practically perpendicular to the PLP's pyridine ring (O1-C $_{\beta}$ -C7-C2 angle: 83.5°). These results agree with Dunathan et al. (69,70), which state that any bound complex formed or cleaved during a reaction involving the PLP cofactor must be placed perpendicularly to the conjugated system of the PLP.

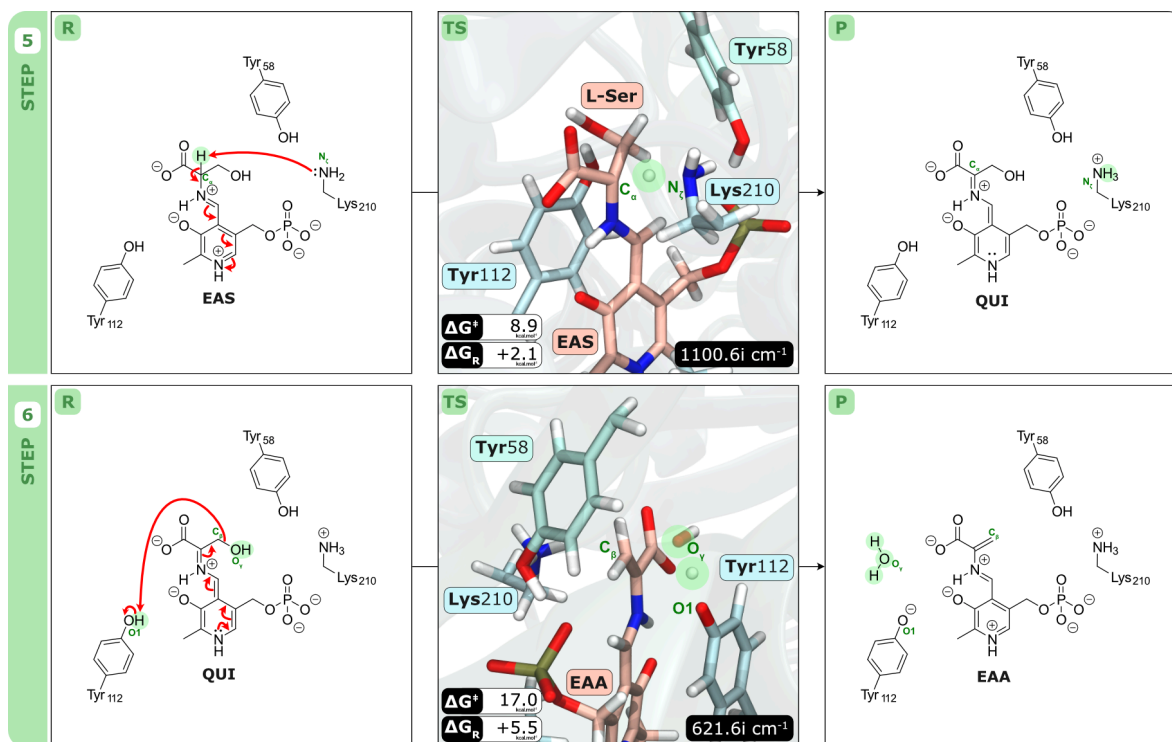


Figure 34. Structures of reactant (R), transition state (TS), and product (P) of steps 5-6. The structures of the reactants and products are represented in a wedge-dash manner. The structures of the TS are represented three-dimensionally, using the Licorice drawing method of VMD. PLP and L-Ser/2AA are represented in pink, while the rest are in light blue. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown.

The formation of the final product of the reaction, pyruvate, has been a question of debate in the literature. It has been proposed by Huang et al. (339) that the 2-aminoacrylate leaves the active site of the enzyme (cystathionine-γ-lyase in this study) still bonded to PLP and that the hydrolysis only occurs in the cytosol. A new molecule of free PLP then binds to the apoenzyme so the holoenzyme can be regenerated. This finding is supported by the fact that when the exogenous concentration of PLP increases, thus the production of H₂S (since the substrate was L-cysteine instead of L-serine). This suggests that the apo-form of the enzyme must form during the catalysis.

Nevertheless, one may argue that the release of the entire external aldimine would require extended alterations to the protein folding, which would take longer than to re-form the internal aldimine and release 2-aminoacrylate from the enzyme's active site. Another proposal was that pyruvate is formed inside the enzyme's active site. Nowadays, however, it is unanimous that 2-aminoacrylate is released to the cytosol, where it is deaminated to pyruvate either non-enzymatically or by another enzyme entitled RidA

(340,341). In order to test this hypothesis, we performed the nucleophilic attack of a water molecule to carbon C_α of the EAA in an attempt to obtain pyruvate. However, the energy barrier of this reaction was $26.6 \text{ kcal.mol}^{-1}$, rendering it impossible on a physiological time scale. According to the transition state theory, this reaction's half-life is 400 hours (i.e., a little over 19 days). These results indicate that the 2-aminoacrylate must leave the active site and go to the cytosol to promote pyruvate formation. These reactions will be described in detail in **section 2.1.3.5**.

2.1.3.3. Stage 3: Formation of the internal aldimine

The release of 2-aminoacrylate from the enzyme requires the cleavage of the covalent bond that it establishes with the PLP cofactor, leading to the formation of the internal aldimine. This is accomplished by the enzyme in seven sequential steps.

In the first step (Step 7), Lys_{210} must lose one of its protons to the phosphate group of PLP, forming a stronger nucleophile (charge of the ϵ -amino group of Lys_{210} : R: 0.41 a. u. vs P: -0.11 a. u.). To do that, Lys_{210} must acquire a better conformation, which is $1.4 \text{ kcal.mol}^{-1}$ more stable than the product of step 6. Only then, thus, the proton transfer occurs with a Gibbs free energy of activation of $4.5 \text{ kcal.mol}^{-1}$ and being endergonic in $1.2 \text{ kcal.mol}^{-1}$ (**Figure 35**). The TS of this reaction is characterized by one imaginary frequency at $857.4i \text{ cm}^{-1}$.

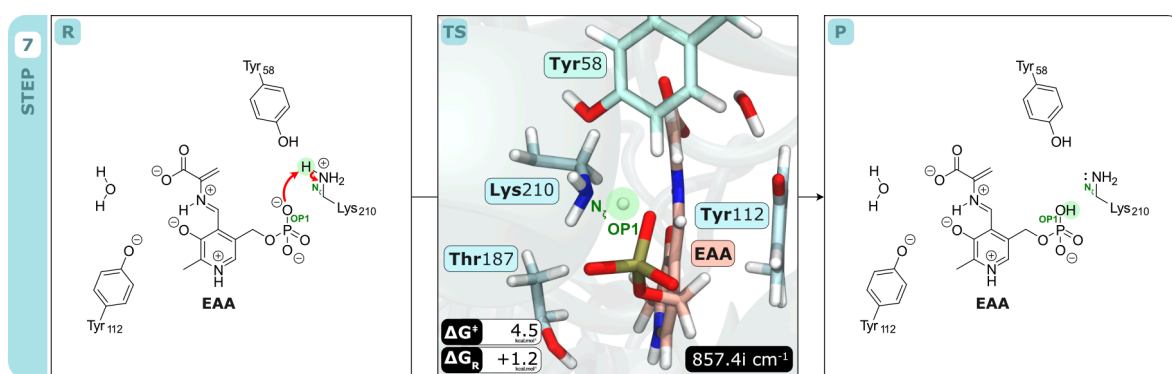


Figure 35. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 7. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. PLP and 2AA are represented in pink, while the rest are in light blue. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown.

In this next step (Step 8), the deprotonated Lys₂₁₀ performs a nucleophilic attack to the carbon C4A of PLP. In this reaction, the bond between this carbon and the nitrogen N1 of 2AA loses its double character (R: 1.31 Å vs P: 1.41 Å), thus forming the gem-diamine reaction intermediate GA1 (**Figure 36**). This reaction, which is basically the reverse of step 4, has a Gibbs free energy of activation of 8.5 kcal.mol⁻¹ and is endergonic in 4.1 kcal.mol⁻¹. The TS structure is characterized by one imaginary frequency at 111.0i cm⁻¹, similar to step 4 (105.2i cm⁻¹) (**Figure 33**).

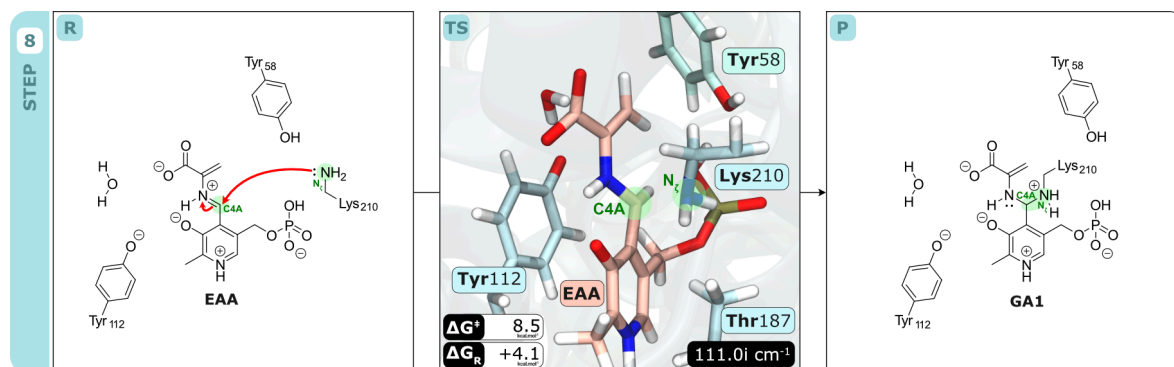


Figure 36. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 8. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. PLP and 2AA are represented in pink, while the rest are in light blue. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown.

The formation of the internal aldimine and release of the 2-aminoacrylate from the gem-diamine GA1 reaction intermediate requires five sequential steps.

First, the proton from the nitrogen N_ζ of Lys₂₁₀ is transferred to oxygen O3 of PLP (**Figure 37** – Step 9). This reaction has a Gibbs free energy of activation of 2.1 kcal.mol⁻¹ and is endergonic in 1.5 kcal.mol⁻¹. Its TS structure is characterized by one imaginary frequency at 948.7i cm⁻¹.

Secondly, the network of hydrogen bonds inside the active site is rearranged. The proton bonded to oxygen O3 of PLP loses its H-bond with nitrogen N_ζ of Lys₂₁₀ and forms a new one with the nitrogen N1 of the 2-aminoacrylate (2AA) (**Figure 37** – Step 10). At the same time, the proton of nitrogen N1 loses its H-bond with the oxygen O3 and now points towards Tyr₁₁₂. This exchange of H-bonds proved to have a Gibbs free energy of activation of 2.3 kcal.mol⁻¹ and is practically thermoneutral, with a Gibbs free energy of reaction of -0.5 kcal.mol⁻¹. Its TS structure is characterized by one imaginary frequency at 124.0i cm⁻¹ involving the movement of both protons.

Thirdly, nitrogen N1 of 2AA abstracts the proton from oxygen O3 of PLP with a Gibbs free energy of activation of $4.2 \text{ kcal.mol}^{-1}$ (**Figure 37** – Step 11). This reaction is endergonic in $4.6 \text{ kcal.mol}^{-1}$ justified by the formation of two point charges in the atoms N1 and O3, respectively. The TS structure of this reaction is characterized by one imaginary frequency at $527.2i \text{ cm}^{-1}$.

Forth, a new rearrangement in the network of H-bonds occurs in an opposite fashion to what was seen in step 2. The proton of nitrogen N_{ζ} that makes one H-bond with Ser₂₀₇ now points to the oxygen O3 of PLP (**Figure 37** – Step 12). This H-bond exchange requires $3.8 \text{ kcal.mol}^{-1}$ and is practically thermoneutral with a Gibbs free energy of reaction of $-0.3 \text{ kcal.mol}^{-1}$. The TS structure of this reaction is characterized by one imaginary frequency at $498.2i \text{ cm}^{-1}$.

Lastly, product 2AA is released, and the internal aldimine is finally formed (**Figure 37** – Step 13). While the bond between the carbon C4A of PLP and nitrogen N1 of 2AA is broken, a double bond between the nitrogen N_{ζ} of Lys₂₁₀ and the C4A of PLP is formed (R: $1.41 \text{ \AA}$ vs P: $1.30 \text{ \AA}$). This reaction has a Gibbs free energy of activation of $4.3 \text{ kcal.mol}^{-1}$ and is exergonic in $2.7 \text{ kcal.mol}^{-1}$. The TS structure of this reaction is characterized by one imaginary frequency at $154.2i \text{ cm}^{-1}$.

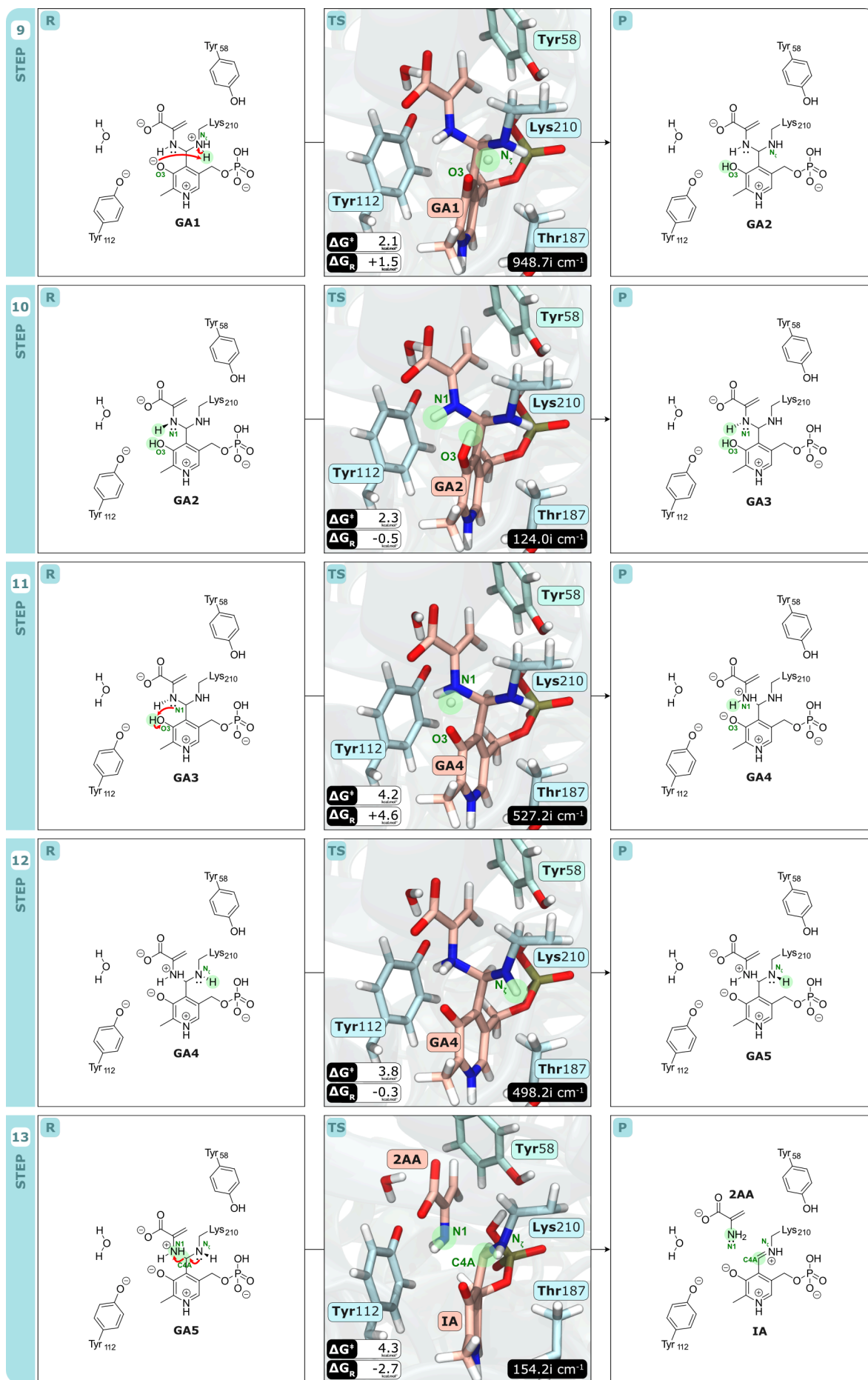


Figure 37. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of steps 9-13. The structures of the reactants and products are represented in a wedge-dash manner. The structures of the TS are represented three-dimensionally, using the Licorice drawing method of VMD. PLP and 2AA are represented in pink, while the rest are in light blue. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown.

2.1.3.4. Stage 4: Reprotonation of Tyr₁₁₂ and enzymatic turnover

The last step in this catalytic mechanism consists of the enzymatic turnover by the reprotonation of Tyr₁₁₂. While the reprotonation of Tyr₁₁₂ could occur after the release of 2AA from the active site, since this residue is exposed to the solvent, we found that the amino group of 2AA could aid in this process by functioning as a proton shuttle between the phosphate group of PLP and the oxygen O1 of Tyr₁₁₂ (**Figure 38**). This final step has a Gibbs free energy of 6.0 kcal.mol⁻¹ and is exergonic in 7.3 kcal.mol⁻¹. Its TS structure is characterized by one imaginary frequency at 791.0i cm⁻¹.

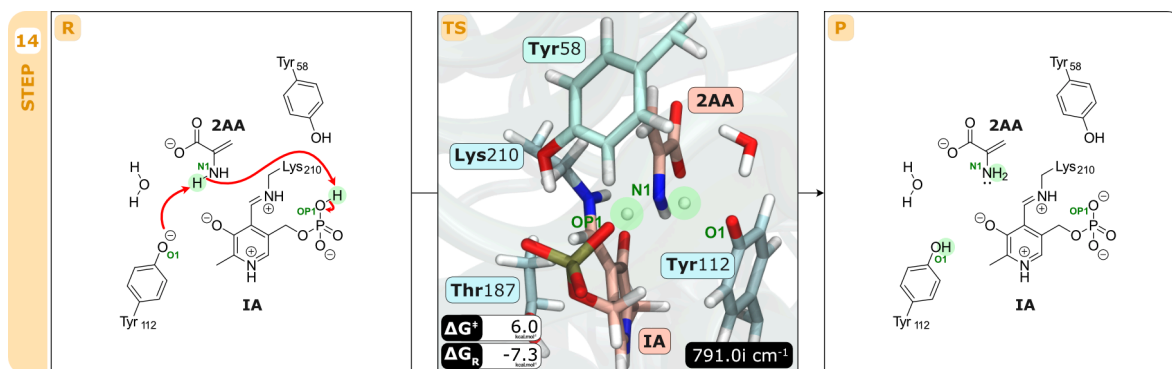


Figure 38. Structures of reactant (**R**), transition state (**TS**), and product (**P**) of step 14. The structures of the reactant and product are represented in a wedge-dash manner. The structure of the TS is represented three-dimensionally, using the Licorice drawing method of VMD. PLP and 2AA are represented in pink, while the rest are in light blue. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown.

2.1.3.5. Non-enzymatic catalysis: Conversion of 2-aminoacrylate into pyruvate

As shown in stage II, the conversion of 2AA into pyruvate must occur in the cytosol, either by non-enzymatic catalysis or by the RidA enzyme. The half-life of 2AA in solution is estimated to be approximately 3 min at 30 °C, pH 7-8 (342). According to the transition

state theory, by using the Eyring equation, this corresponds to an energy barrier of about 21 kcal.mol⁻¹, which must be the rate-limiting step of this reaction.

The deamination of the 2AA to pyruvate starts with the tautomerization of the enamine form into 2-iminopropionate, its imine form (**Figure 39** – Step 1). This reaction requires two water molecules that work as a proton shuttle between the nitrogen N1 and the carbon C_β. In this process, the bond between the carbons C_α and C_β loses its double-bond character (R: 1.35 Å vs P: 1.49 Å), whereas a new double-bond is formed between carbon C_α and nitrogen N1 (R: 1.38 Å vs P: 1.28 Å). This reaction has a Gibbs free energy of activation of 22.6 kcal.mol⁻¹ and is exergonic in 6.4 kcal.mol⁻¹. The TS structure of this reaction is characterized by one imaginary frequency at 435.0i cm⁻¹.

In the next step, the water molecule WAT3 attacks the carbon C_α of 2-iminopropionate, while two other water molecules shuttle a proton from this water to nitrogen N1, making its bond with the carbon C_α lose its double-bond character (R: 1.28 Å vs P: 1.43 Å) (**Figure 39** – Step 2). This reaction has a Gibbs free energy of activation of 21.2 kcal.mol⁻¹ and is endergonic in 3.8 kcal.mol⁻¹. This TS is characterized by one imaginary frequency at 639.0i cm⁻¹.

In the third step (**Figure 39**), the oxygen OW3 of the newly added hydroxyl group loses its proton to nitrogen N1 of the amino group. Two water molecules mediate this proton transfer. This reaction requires a Gibbs free energy of activation of 10.8 kcal.mol⁻¹ and is endergonic in 8.4 kcal.mol⁻¹, justified by the formation of two point charges in the molecule. The TS obtained for this reaction is characterized by one imaginary frequency at 659.2i cm⁻¹.

Lastly, the carbonyl double bond is formed between carbon C_α and oxygen OW3 (R: 1.34 Å vs P: 1.22 Å) with the simultaneous elimination of the ammonia molecule (**Figure 39** – Step 4). This reaction requires a Gibbs free energy of activation of 4.7 kcal.mol⁻¹ and is exergonic in 4.8 kcal.mol⁻¹. The TS structure is characterized by an imaginary frequency at 184.9i cm⁻¹.

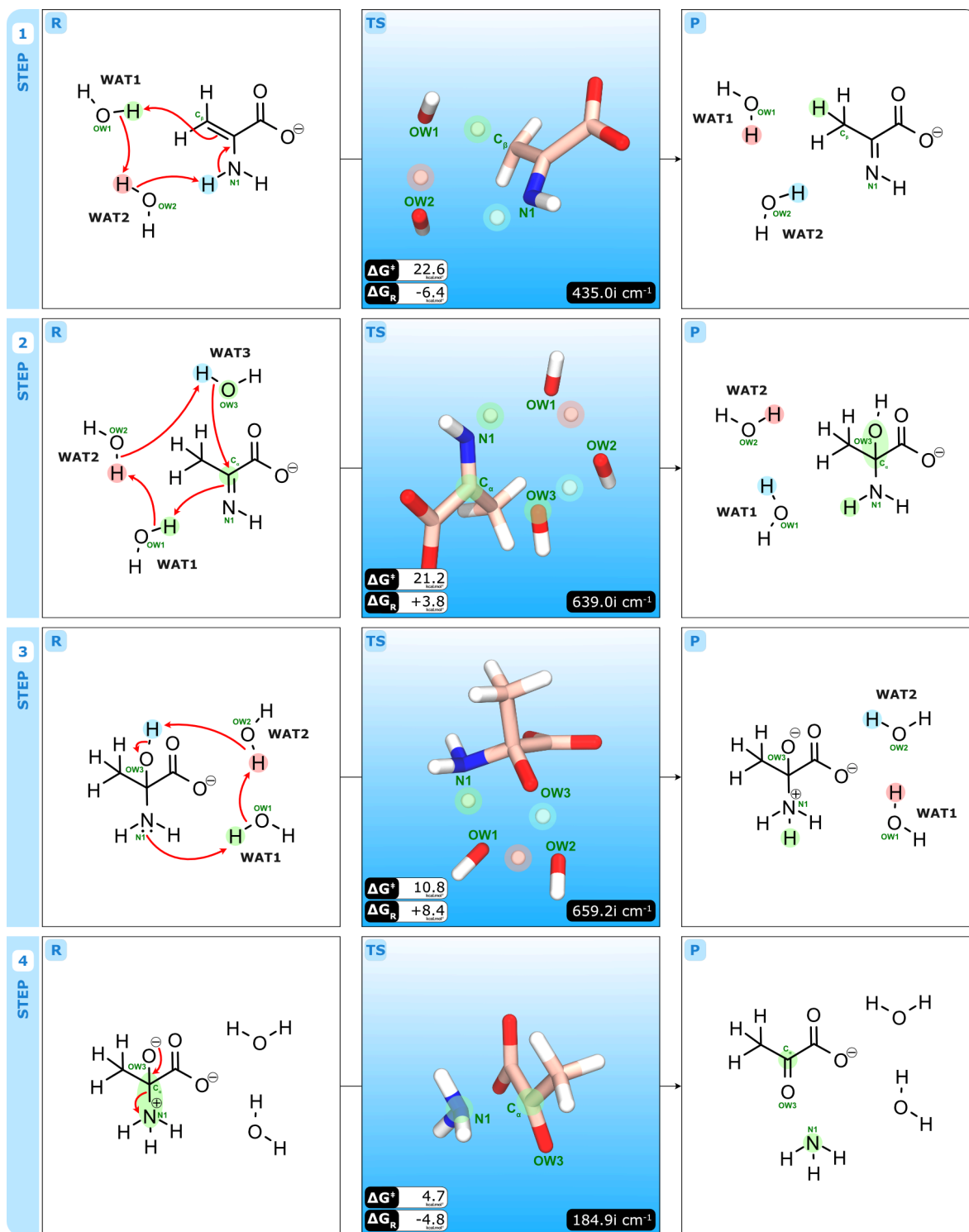


Figure 39. Structures of reactant (R), transition state (TS), and product (P) of the steps involved in the non-enzymatic catalysis of the deamination of 2AA into pyruvate that occurs in the cytosol. The structures of the reactants and products are represented in a wedge-dash manner. The structures of the TS are represented three-dimensionally, using the Licorice drawing method of VMD, and are coloured in pink. The imaginary frequency corresponding to the major vibrational modes and the activation and reaction Gibbs free energies are shown.

2.1.3.6. UV-VIS

TD-DFT calculations were carried out to compare the absorption spectra of the intermediates of XometC's catalytic mechanism with the ones experimentally determined for PLP-dependent enzymes. This allows us to corroborate the intermediary structures computationally obtained in this work. Absorption spectra (**Figure 40**) were calculated for four intermediates in this catalytic mechanism: internal aldimine (IA), external aldimine with L-serine (EAS), quinonoid (QUI), and the external aldimine with 2-aminoacrylate (EAA).

It is known that these stable PLP reaction intermediates have characteristic absorption bands that can be detected experimentally. The IA has a characteristic UV-Vis absorption peak in the range of 410-420 nm (343–347), EAS in the range of 400-435 nm (343,345,348,349), QUI around 500 nm (343,344,350), and EAA around 460 nm (343,345).

The *in silico* absorption maxima obtained for each of these structures were, by order of appearance in the catalytic mechanism, 383 nm for the IA, 404 nm for the EAS, 437 nm for the QUI, and 431 nm for the EAA.

Although the calculated absorption bands are shifted relative to the experimental values, they show a similar trend. The quinonoid intermediate is the structure with the absorption peak at the highest wavelength, followed by EAA, EAS, and IA. These results align with experimental observations that show that extended conjugated pi systems tend to have UV-Vis bands at longer wavelengths. This is the case of the quinonoid intermediate, where the PLP cofactor is negatively charged, and the charge is distributed along the cofactor and part of the L-Ser. The external aldimine band of EAA and EAS appears at higher wavelengths than IA because they have more double bonds, some conjugated, which increases the absorption wavelength.

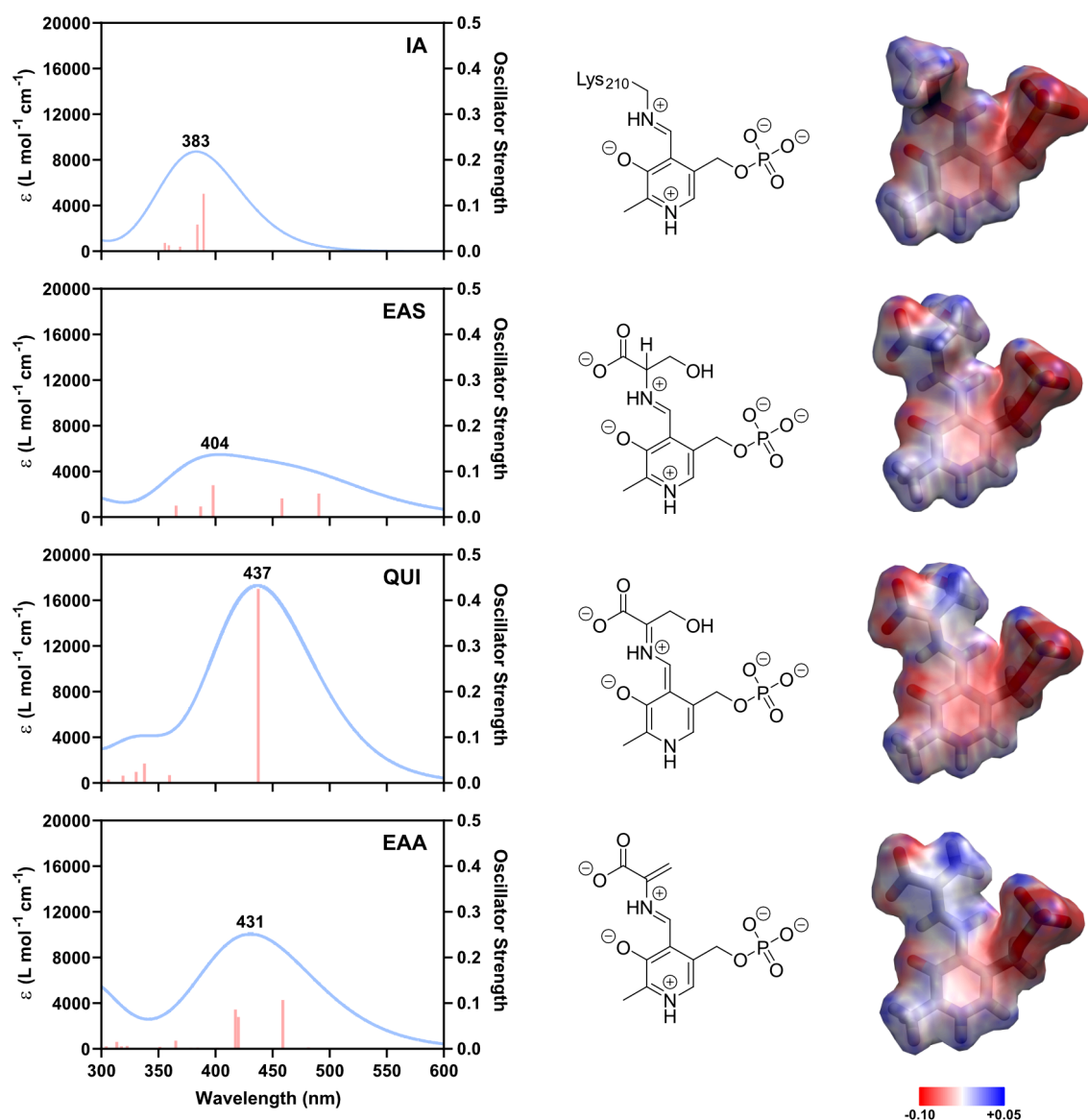


Figure 40. UV-Vis absorption spectra calculated with TD-DFT of four important intermediates of the catalytic mechanism of XometC: the internal aldimine (IA), the external aldimine with L-serine (EAS), the quinonoid (QUI), and the external aldimine with 2-aminoacrylate (EAA). The electrostatic potential was calculated and mapped on the surface of the reaction intermediates.

2.1.4. Conclusion

In this work, the catalytic mechanism of the XometC enzyme was studied with an atomistic level of detail using computational QM/MM methodologies. XometC was initially described as a cystathionine γ -lyase, but it was found that it also presents L-serine ammonia-lyase activity, which was the main focus of this work. This reaction involves the conversion of L-serine into 2-aminoacrylate with the aid of the PLP cofactor. The reaction

consists of a series of 14 sequential steps that can be divided into four main stages: the formation of the external aldimine with L-serine, the β -elimination, the re-formation of the internal aldimine and release of 2-aminoacrylate, and finally, the enzymatic turnover with the reprotonation of Tyr₁₁₂.

The complete mechanism is depicted in **Figure 41**. In the first step, L-Ser attacks the carbon C4A of the internal aldimine, forming a gem-diamine reaction intermediate. In the second step, the H-bond performed by the nitrogen N $_{\zeta}$ of Lys₂₁₀ changes from the oxygen O3 of PLP to Ser₂₀₇, allowing the nitrogen N1 of L-Ser to perform a stronger H-bond with oxygen O3. Then, in steps 3 and 4, the proton from the nitrogen N1 of L-Ser that was establishing the H-bond is spontaneously transferred to the oxygen O3 of PLP and from there to the nitrogen N $_{\zeta}$ of Lys₂₁₀. Finally, the external aldimine was formed, and the neutral Lys₂₁₀ was released. In the β -elimination stage, Lys₂₁₀ abstracts the proton from the carbon C $_{\alpha}$ of L-Ser (Step 5), and the elimination of the hydroxyl group occurs in the form of a water molecule, with the second proton coming from Tyr₁₁₂ (Step 6). After this stage, the product, 2-aminoacrylate, is formed.

The possibility that the β -elimination stage could occur concertedly was also considered. In this scenario, the proton from the carbon C $_{\alpha}$ of L-Ser would be abstracted by the nitrogen N $_{\zeta}$ of Lys₂₁₀ with the concomitant elimination of the hydroxyl group of L-Ser. This would mean that PLP would only function as an anchor for L-Ser and that no quinonoid would be formed. This is the mechanism employed by other L-serine ammonia-lyases because since their pyridine ring is not protonated, PLP cannot function as an electron sink, stabilizing the charge in its ring (310–314). However, this hypothesis was discarded since we found it impossible to occur in a physiological timescale, requiring over 40 kcal.mol⁻¹. Therefore, the β -elimination must occur in two steps, as shown.

However, 2-aminoacrylate remains covalently bonded to PLP as an external aldimine. Therefore, to release it, the reverse of stage I must occur. After Lys₂₁₀ loses one of its protons to the phosphate group (Step 7), it attacks the carbon C4A of PLP to form a new gem-diamine (Step 8). From there, one of the protons at the nitrogen N $_{\zeta}$ of Lys₂₁₀ is transferred to the oxygen O3 of PLP (Step 9). This makes the proton at the nitrogen N1 of 2AA lose its H-bond with oxygen O3 (Step 10), allowing the proton from the oxygen to be transferred to the nitrogen N1 (Step 11). Then, similarly to step 2, the proton of the nitrogen N1 of Lys₂₁₀ exchanges H-bonds (Step 12), and the internal aldimine is formed with the release of 2AA (Step 13). In the end, the reprotonation of Tyr₁₁₂ occurs with the help of the newly formed 2AA (Step 14), which can now leave the active site, making the enzyme ready for a new enzymatic cycle.

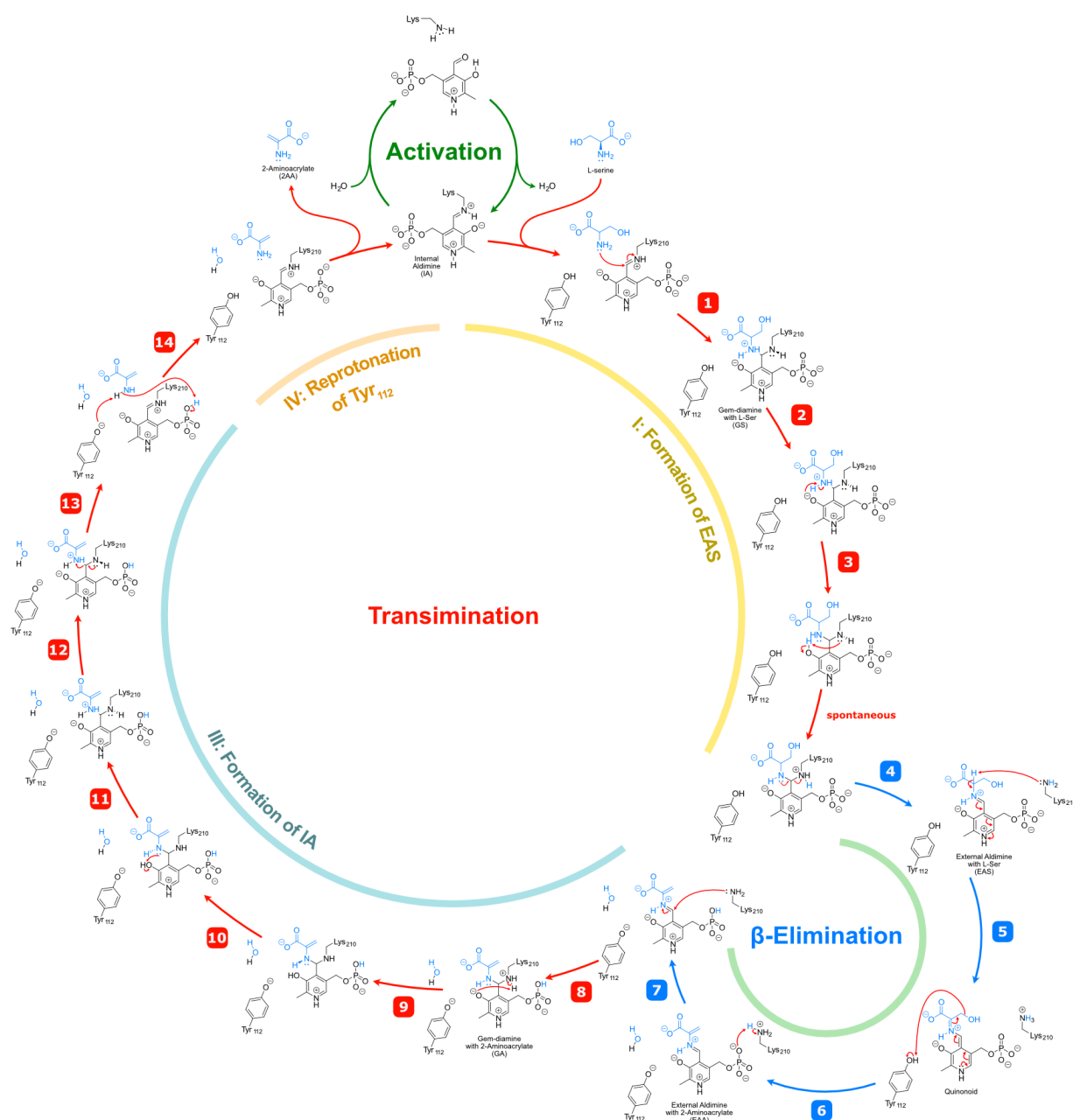


Figure 41. Overview of the catalytic mechanism of the PLP-dependent enzyme XometC uncovered in this work.

According to the determined free energy profile depicted in **Figure 42**, the proton abstraction by the Lys₂₁₀ (Step 5) and the elimination of the hydroxyl group of L-serine (Step 6) contribute synergically to the rate-limiting step of the catalytic process, with a total activation free energy barrier of 19.1 kcal.mol⁻¹. These results closely agree with the experimental kinetic data available, whose activation barrier predicted is 18.9 kcal.mol⁻¹ (72).

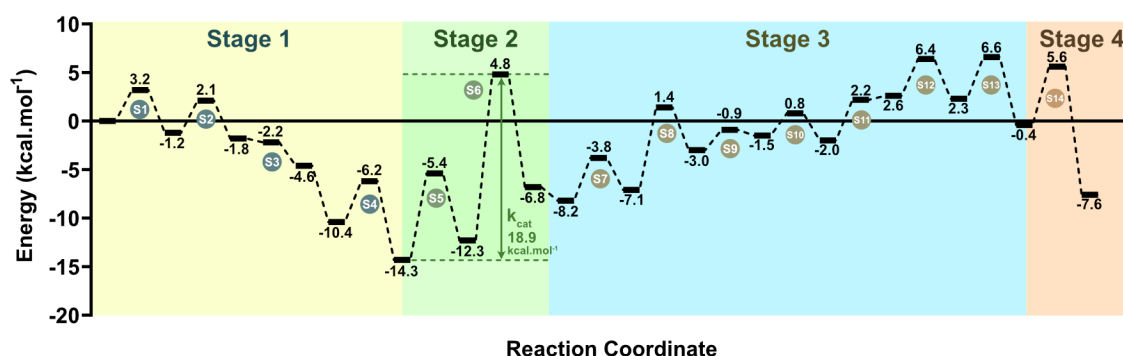


Figure 42. Energy profile determined for the catalysis of L-serine conversion into 2-aminoacrylate by XometC. All energy values were computed with DLPNO-CCSD(T)/CBS//r²SCAN-3c/mTZVPP:ff14SB.

It is worth noting that the free energy profile presents a valley shape, with its energy decreasing until the formation of the external aldimine, which then increases once again until the release of 2-aminoacrylate. This observation is indicative that one of the major roles of the enzyme is to form and stabilize the external aldimine. This is important because only in this reaction intermediate is the substrate placed perpendicularly to the conjugated system of PLP, favouring by this way the cleavage of a covalent bond, as proposed by Dunathan (69,70) and verified in this work. Generally, the cleavage of covalent bonds is difficult to occur under physiological conditions because they require high activation energies. However, this work shows that this reaction is possible under physiological conditions due to the involvement of the PLP cofactor and the correct alignment of the substrate endorsed by the active site residues strategically positioned by the enzyme to favour the chemical reaction.

Moreover, we have shown that the conversion of 2-aminoacrylate into pyruvate cannot occur under enzymatic catalysis since it requires high activation energies (above 25.0 kcal.mol⁻¹). Instead, this conversion occurs in the cytosol, with a rate-limiting step (i.e., Step 1 – tautomerization) whose activation barrier (22.6 kcal.mol⁻¹; **Figure 43**) closely agrees with the experimental data described in the literature (21 kcal.mol⁻¹) (342). This energy value corresponds to a reaction half-life of approximately 3 minutes, which is a reasonable value at a physiological timescale. However, since the 2-aminoacrylate is a very toxic molecule with a high half-life that can irreversibly inhibit PLP-dependent enzymes, cells have other enzymes, such as 2-iminobutanoate/2-iminopropanoate deaminase (also known as RidA), which accelerate this deamination process. This efficiently reduces the potential cellular damage that can be caused by 2-aminoacrylate (340,341,351–353).

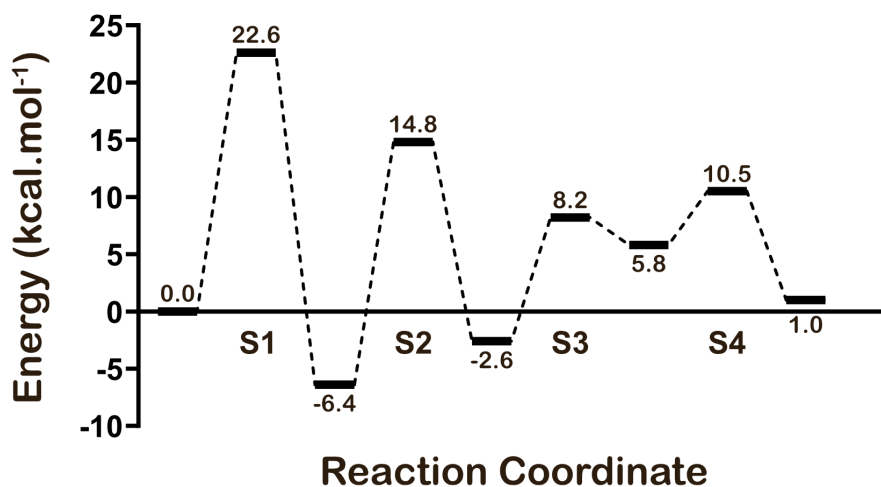


Figure 43. Energy profile determined for the non-enzymatic catalysis of the deamination of 2-aminoacrylate to pyruvate. All energy values were computed with DLPNO-CCSD(T)/CBS.

Together, the results obtained in this work provide a complete overview, with an atomistic level of detail, of the catalytic mechanism of XometC that can aid in developing new drugs against the bacterial blight of rice. Particularly, transition state analogue inhibitors can be designed from the transition state of step 5, leading to a highly stabilized intermediate that could stall the entire catalytic mechanism. Furthermore, since the target enzyme is not present in humans, the inhibitors would be highly specific and have low toxic effects for humans.

This page was intentionally left blank.

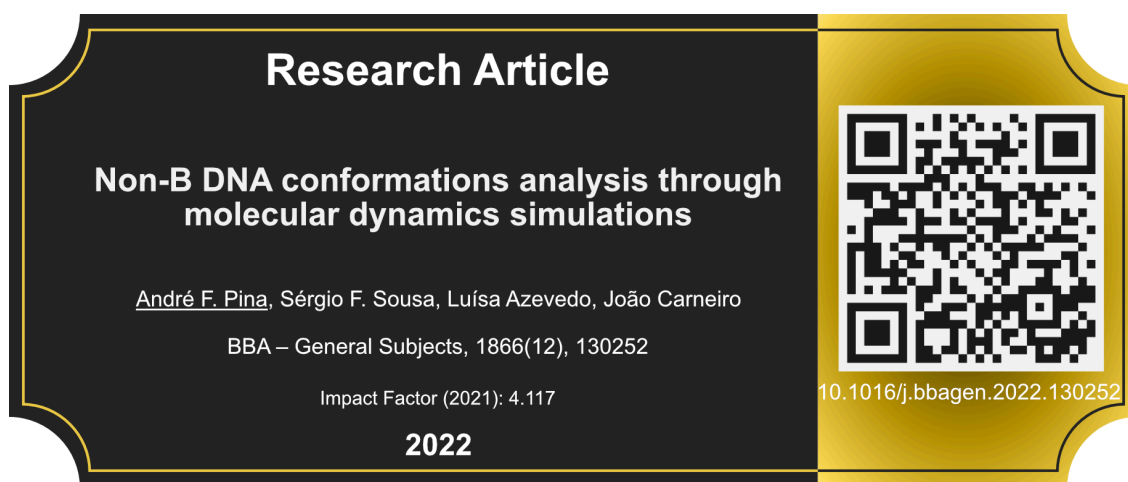
3. Non-B DNA conformations

This page intentionally left blank.

3.1. Analysis of the 3D conformation of Structure A – a non-B DNA conformation associated with mtDNA mutagenic instability

Non-B DNA conformations are molecular structures that do not follow the canonical DNA double helix. Mutagenetic instability in nuclear and mitochondrial DNA (mtDNA) genomes has been associated with simple non-B DNA conformations, such as hairpins, or more complex structures, such as G-quadruplexes. One of these structures is Structure A, a cloverleaf-like non-B conformation predicted for a 93-nt (nucleotide) stretch of the mtDNA control region 5'-peripheral domain. Furthermore, Structure A is embedded in a hot spot for the 3' end of human mtDNA deletions revealing its importance in influencing the mutational instability of the mtDNA genome. In order to characterize Structure A better, its 3D conformation was predicted using state-of-art methods and algorithms. Then, molecular dynamics simulations were performed to study its stability.

This work resulted from a strict collaboration with João Carneiro from CIIMAR, Porto, and Luísa Azevedo from i3S, Porto. André F. Pina performed the 3D predictions of all non-B DNA conformations under study, their molecular dynamics simulations, and subsequent analyses. Regarding the article, André F. Pina wrote all the methodology, results, and discussion related to the work mentioned above, having also partially contributed to the writing of the introduction. The following content is almost a transcript of the published version, which was published in the *Biochimica et Biophysica Acta (BBA) – General Subjects* journal.



In short, Structure A 3D conformation showed high stability and low flexibility, which was correlated with high conservation scores in mammals, specifically primate lineages. Furthermore, the results showed that 3D non-B conformations can be predicted and

characterized with the methodology used, and their stability can be estimated using the base pairing level maintained throughout the simulation.

This work was also presented at the following conference:

1. (Poster) *Bioinformatics Open Days*, **2021**, online (**Appendix:** page 264)

3.1.1. Introduction

Non-B DNA conformations are molecules whose conformations do not follow the double helix DNA structure. They can assume simple conformations as DNA hairpins or more complex structures as G-quadruplexes. The DNA's structure of these conformations is related to its base sequence and to the solvent environment where it is located. Since DNA is a highly charged polyanion, cations have an important role in its stabilisation, such as reducing phosphate-phosphate repulsion. Some studies have shown that different ions with different ionic sizes have a different impact on the stability of DNA (354–358). However, long-timescale molecular dynamics simulations have shown that DNA conformations are insensitive to the ionic strength of the solution or the nature of the salt (359). Nevertheless, there is more than just water and ions inside a cell. The entire volume of a cell is occupied by a large number of biomolecules, which is the reason why it is said that the cellular environment is crowded, a phenomenon referred to as macromolecular crowding (360–362). How molecular crowders affect nucleic acids' structure, interactions, and function has been a matter of great study. It has been shown that molecular crowding plays a critical role in determining the structure and stability of G-quadruplex DNA (363,364). The same has also been verified for G-triplexes (365).

These conformations do not follow a random distribution in the human nuclear genome, and several can have functional implications for cellular functions and human diseases (85). In this context, non-B conformations can fold in approximately 13% of the human nuclear genome and can determine natural obstructions to the DNA synthesis elongation by the replicative polymerases (85,366,367). In human mitochondrial genomes, these structural conformations can also occur (89,95). A complete analysis of the occurrence of non-B conformations in human mitochondrial DNA (mtDNA) has been done by Pereira et al., which demonstrated their relation with the rate of nucleotide substitutions and with the 3' end of human mtDNA deletions (89). Such DNA conformations are highly related to events of mutagenetic instabilities in both nuclear and mitochondrial DNA (86–88,95,368–375) and, therefore, are associated with several diseases that result from replication errors, repair errors, and genomic instability (87,95,376–378). Most replication errors can be associated with mtDNA deletion hotspots where non-B conformations can be formed (95). The factors that influence the mtDNA non-B conformations formation have been under investigation using both a somatic (within species) and an evolutionary (inter-species) approach (95,370). In both frameworks, free energy calculations of different regions of the mtDNA have shown that regions with high folding potential were statistically correlated with genomic rearrangements implied in mtDNA breakage hotspots. One example of a non-B conformation that influences the replicative process in the

mitochondria is Structure A, a cruciform conformation formed near the origin of the replication region (89). The formation of non-B conformations like Structure A in the mtDNA control region results from the occurrence of extensive single-stranded DNA stretches during mtDNA replication or in the three-stranded D-loop (95), which can constitute pausing sites for the polymerase activity, inducing an obstacle to fork progression. The study of Structure A is relevant to understand processes that originate mtDNA deleted molecules as a result of hypermutable non-B DNA.

The Unified Nucleic Acid Folding and hybridisation package (UNAFold) is a well-established software that uses experimentally determined thermodynamic parameters to predict non-canonical RNA and DNA secondary structures (379). The free energy calculations can improve the computational methods to build molecular models of different non-B conformation DNA systems. In this way, computational approaches, such as molecular dynamics (MD), have been used to better understand the behaviour of structural DNA elements in different environments (380–391). Different types of non-B DNA, including the DNA hairpin, pseudoknot, and cruciform conformations, can be evaluated using stability parameters during MD simulations.

We performed MD simulations to mimic the DNA conformations in Structure A, considering the UNAFold secondary structure predictions and their putative implications in replication mechanisms (95,377). Our innovative approach was designed to explain the Structure A stability at the structural level and the frequency at which it can be formed in the mtDNA. Combining these data with evolutionary conservation scores (392–394) using lineage-specific genome alignments for primates and other mammals demonstrates that structure A conformation can be maintained during the molecular dynamics simulations. The results also suggest that the nucleotide composition significantly influences the maintenance of non-B conformations in mammals and primates.

3.1.2. Methodology

3.1.2.1. Determination of secondary structures

Structure A is located between positions 16,028–16,120 of the mtDNA (according to the reference sequence NC_012920) at the hypervariable region I (HVR-I) and D-loop 3' end. The minimum free energy (MFE) of the Structure A (DNA sequence - 5' [TTCATGGGGAAGCAGATTTGGGTACCAACCAAGTATTGACTCACCCATCAACAACCG CTATGTATTTTCGTACATTACTGCCAGCCACCATGAA] 3') was calculated using the

latest online version of the UNAFold algorithm that uses free energy minimisation using nearest neighbour thermodynamic rules (379,395). The minimum free energy (MFE) values were also calculated for the control sequences: a) 93-nt single-stranded nucleotide sequence of Structure A; b) 93-nt mono-shuffle nucleotide sequence of Structure A; c) 93-nt 5' and 3' Structure A flanking sequences; d) lowest and highest MFE structures from a dataset of all 93-nt windows of the complete human mtDNA genome. The single-stranded conformation was used as the baseline for comparison with the non-B conformations (controls and Structure A). The mono-shuffle was used to measure the influence of nucleotide composition on non-B conformation maintenance. The flanking regions were used to show if the adjacent regions of Structure A present the same values of MFE and similar parameters (e.g., h-bonds). The lowest and highest MFE were used to ascertain the range of MFE, which includes upper and down limits, of non-B conformation that occur in mtDNA. We used the UNAFold software to predict the secondary structure using the following parameters: temperature of 37 °C and ions concentration between the minimum and maximum concentrations used in PCR experiments (sodium [Na⁺]: 0.5 and 1 mol.L⁻¹, magnesium [Mg²⁺] 0.002 and 0.005 (mol.L⁻¹), and “window size parameter” of 2. All folding free energies are expressed in kilocalories per mole (kcal.mol⁻¹). Temperature and ion conditions were selected to mimic a standard PCR system at the starting temperature (396–399). The tested concentration revealed no variation in the base pairing of the predicted final structure. The consensus base pairing between all calculations was used. We used the PCR parameters as a simplified model representing the most important variables of the DNA single-stranded molecules in *in vivo* cellular processes.

3.1.2.2. UNAFold predictions accuracy

The accuracy of the Unified Nucleic Acid Folding and hybridisation package – UNAFold – was assessed using a database of 22 experimentally-determined nuclear magnetic resonance (NMR) spectroscopy single-stranded DNA structures (**Table 1**) with hairpins and stems. These 3D structures were compared with the secondary structure prediction of UNAFold.

Table 1. Accuracy results of UNAFold to predict hairpin structures using a database of experimentally determined NMR structures of 22 single-stranded DNA with hairpins.

PDB ID	3D structure type	UNAFold prediction (base pairing and loop)	Length (nt)	Percentage of base pair
1AC7	ssDNA hairpin	Identical to 3D	16	100
1BJH	ssDNA hairpin	Identical to 3D	11	100
1DGO	ssDNA hairpin	Identical to 3D	18	100
1ECU	ssDNA hairpin	Identical to 3D	19	100
1EL2	DNA complex telomeric structure (3 loops)	Different from 3D	22	100
1ELN	DNA complex telomeric structure (3 loops)	Different from 3D	22	100
1EN1	Primer binding site HIV - ssDNA hairpin	Identical to 3D	18	100
1FV8	ssDNA heterochiral hairpin	Identical to 3D	11	100
1IDX	ssDNA hairpin	Identical to 3D	18	100
1III	ssDNA hairpin	Identical to 3D	18	100
1JVE	AT-Rich ssDNA with the GAA-Hairpin Loop	Identical to 3D	27	100
1KR8	ssDNA hairpin	Identical to 3D	7	100
1LA8	ssDNA hairpin	Identical to 3D	13	100
1LAE	ssDNA hairpin	Identical to 3D	13	100
1NGO	ssDNA hairpin	Identical to 3D	27	100
1NGU	ssDNA hairpin	Identical to 3D (except 2 residues in stem)	27	90
1P0U	ssDNA triplet repeats hairpin	Identical to 3D	13	100
1PQT	ssDNA hairpin	Identical to 3D	7	100
1QE7	ssDNA hairpin	Identical to 3D	22	100
1XUE	ssDNA hairpin	Identical to 3D only for loop	17	29
1ZHU	Hairpin loop formed by the DNA triplet GCA.	Identical to 3D	10	100
2M22	ssDNA hairpin	Identical to 3D	23	100

3.1.2.3. Generation of three-dimensional non-B DNA conformations

The generation of the three-dimensional (3D) non-B DNA conformations occurred in two steps:

Step 1: Generation of equivalent 3D single-stranded (ss) RNA models with RNAComposer

The secondary structure predicted by the UNAFold software was used to generate 3D models of the Non-B DNA conformations using the RNAComposer web server (400,401). Due to the lack of available tools for predicting 3D ssDNA structures (402), an equivalent 3D RNA structure was generated. This approach has been previously employed and validated by the works of Heiat et al., 2016, Jeddi & Saiz, 2017, and Sabri et al., 2020 (402–404). The predictions from UNAFold were used as input in the dbn file format. No secondary structure prediction method was selected since it is already provided. The resulting structures were then downloaded in the pdb file format.

Step 2: Conversion of the generated 3D ssRNA models into ssDNA using LEaP from the AmberTools software

Lastly, the 3D RNA structures were converted to DNA. First, an in-house bash script was used to remove the hydroxyl group at the C2' position of the ribose ring and the hydrogen atom at the C5 position of uracil. Then, the LEaP program, part of the AmberTools software package (283), was used to add all the missing atoms, including the methyl group needed to convert uracil into thymine. The resulting structures are now 3D ssDNA molecules.

3.1.2.4. Molecular dynamics simulations of the 3D non-B DNA conformations

Molecular dynamics simulations were performed considering a total of seven structures: six control structures (i.e., 93-nt linear single-stranded Structure A, 93-nt Structure A after mono-nucleotide shuffle sequence, two 93-nt sequences with the highest and lowest folding potentials in the complete mtDNA genome, and the two 93-nt flanking sequences of the Structure A region), and the 93-nt Structure A.

These non-B DNA conformations were solvated with TIP3P (286) water molecules in a rectangular box so that each atom of the DNA was distanced at least 12 Å from the edge of the periodic box. Then, a physiological concentration of NaCl (150 mM) was inserted into the system by the LeaP package of the AmberTools18 software (283). The number of added Na⁺ and Cl⁻ ions was determined using the Screening Layer Tally by Container Average Potential (SLTCAP) method (405). The ions were added in two steps. Firstly, enough Na⁺ ions were added to neutralise the DNA molecule, and then an equal number

of Na^+ and Cl^- ions were placed randomly to obtain the desired salt concentration. The ssDNA molecules were described by the bsc1 force field, developed for the atomistic simulation of nucleic acids (406).

Four steps of energy minimisation were performed to reduce possible constraints and clashes. Firstly, only the water molecules were minimised (2500 steps). Then, all hydrogen atoms were minimised, followed by all non-backbone atoms (500 steps each). Lastly, all atoms in the system were freely minimised with no constraints (10,000 steps). Information about the evolution of the system throughout each step of the minimisation can be found in the supplementary material (**Figs. S1 to S3**). Next, the system equilibration was performed in a two steps process. Firstly, the system was gradually heated from 0 to 310.15 K, over 20 ps applying weak restraints on DNA. Then, the restraints were removed, and the system was maintained at this temperature for another 80 ps.

The production phase of the MD simulations was run for 50,000,000 steps, with an integration time of 2.0 fs, totalling 100 ns. Three replicates were performed in two representative systems to evaluate the impact of more replicates in the analysis of the results, which can be found in Figs. S4 and S5. The SHAKE (139) algorithm was used to constrain all bonds involving hydrogen atoms. The system was maintained at constant temperature (310.15 K; friction coefficient of 1 ps^{-1}) and pressure (1.0 bar), using the Langevin thermostat (407) and Berendsen barostat (146). A side study was performed where we evaluated the validity of the use of the Berendsen barostat in this work. The results obtained can be found in **Fig. S6** of the supplementary material. A cut-off of 10 Å was applied to the non-bonded interactions. Both equilibration and production phases were conducted using the PMEMD software from the AMBER18 package.

3.1.2.5. Analysis of the molecular dynamics simulations

The following properties of the molecular dynamics simulations for each non-B DNA conformation were analysed: root-mean-square deviation (RMSD), radius of gyration (R_g), end-to-end distance (R_{ee}), and the number of intramolecular hydrogen bonds. Their analysis was performed using CPPTRAJ, a program distributed within the AmberTools software package for processing and analysing trajectory data from molecular dynamics simulations.

3.1.2.6. Root-Mean-Square Deviation (RMSD), end-to-end Distance (R_{ee}), and radius of gyration (R_g)

We analysed the MD stability by calculating the mass-weighted RMSD using all heavy atoms in the system. The distance between the centres of mass of the 3' and 5' end residues of the non-B DNA conformations was also determined and plotted as a function of time for the 100 ns of the simulation. The mass-weighted radius of gyration of all atoms in the system was also calculated and plotted as a function of time for the 100 ns of the simulation.

3.1.2.7. Hairpin states and H-bond trajectory analysis

The process of hairpin formation described in Portella et al. (408) was used to analyse the different folding mechanisms. The hairpin states in each molecular model were analysed by considering three regions of low MFE: 1) the native form (with 6–8 hydrogen bonds and around four stacked bases); 2) a partially folded state characterized by two hydrogen bonds and two stacked bases, and 3) the fully unfolded form with no native hydrogen bonds and a variable number of stacked bases (408). This process involving loop nucleation with one base pairing connection followed by second base pairing connection results from fluctuations between closed and open hairpin states (384,408–410).

Intramolecular hydrogen bonds formed between base pairs throughout the trajectory were analysed with CPPTRAJ. The hydrogen bond donors and acceptors were defined via the following atom masks: DA@N6, DT@N3, DG@N1, DG@N2, and DC@N4 as the donors; DA@N1, DT@O4, DG@O6, DC@N3, DC@O2 as the acceptors. A distance of 3.0 Å from the donor to the acceptor heavy atoms was used as a cut-off.

3.1.2.8. Cluster analysis

A cluster analysis of the conformations adopted by each non-B DNA sequence during the molecular dynamics simulation was performed using the K-means clustering algorithm. A maximum of 5 clusters were defined for each trajectory. The RMSD of all heavy atoms was used as a measure of similarity (i.e., the smaller the RMSD, the more similar the structures) to form the clusters. A representative structure for each of the five main clusters for each non-B DNA conformation was saved in PDB format.

3.1.2.9. Conservation scores analysis

We have performed an in-deep analysis of the evolutionary conservation of Structure A across humans, non-human primates, and other mammals.

We used a reference alignment of 44,299 mtDNA complete genomes retrieved from the work of McInerney and colleagues (411) to analyse the conservation in humans. We also evaluated the mammal's and primate's inter-species nucleotide conservation using the complete mtDNA genomes from NCBI (November 2021). The mammalian and primate sequence alignments were obtained using MAFFT (412) at <https://usegalaxy.eu/> with default parameters. Geneious software (413) was used to determine the Structure A localisation in all alignments. Structure A is located between positions 16,028 and 16,120 in the human genomes and between 17,106–17,242 and 2684–28,732 in primate and mammalian alignments, respectively. The Structure A region was analysed considering the following conservation scores for each nucleotide position (window size of 1 and step of 1): percentage of identical sites (PIS) and percentage of pairwise identity (PPI) (394). The disease-associated variants and somatic mutations were retrieved from Mitomap (414). The correlation between the mean RMSF of each molecular dynamics, conservation scores, and the disease-associated variants was calculated in JASP (415).

3.1.3. Results and Discussion

3.1.3.1. UNAFold predictions accuracy

Considering the NMR 3D structures analysed to evaluate the UNAFold accuracy, we obtained an overall result that gives high performance for the hairpin prediction of UNAFold software. We have used a database of experimentally determined NMR structures of 22 single-stranded DNA with hairpins (**Table 1**, <http://ndbserver.rutgers.edu/index.html>) to evaluate the accuracy of UNAFold in predicting hairpin structures. We obtained 20 accurate loop predictions (90.9%) and 18 accurate stem and loop predictions out of 22 (81.8%). The structures that UNAFold could not predict were PDB 1EL2 and PDB 1ELN because these are complex 3D structures that fold over themselves, making them unreliable for UNAFold prediction.

3.1.3.2. Structure A is stable throughout molecular dynamics simulations

The structures were analysed using the mean of RMSD between the initial UNAFold predicted structure and the time of the stabilisation of the structure as criteria. The analysis of the molecular dynamics simulations showed a highly stable Structure A, with a mean RMSD of 10 Å, when compared with the initial UNAFold predicted structure. The controls always presented a mean RMSD value much higher than the initial UNAFold prediction (**Figure 44, Table 2**). The single-stranded control was used as a baseline to measure the RMSD of the Structure A region when Structure A is not formed due to mtDNA cell conditions. When comparing the single-strand with the other non-B conformations controls, the lower number of h-bonds formed shows a considerable difference from all other conformations, including the structure with the highest MFE in the mtDNA.

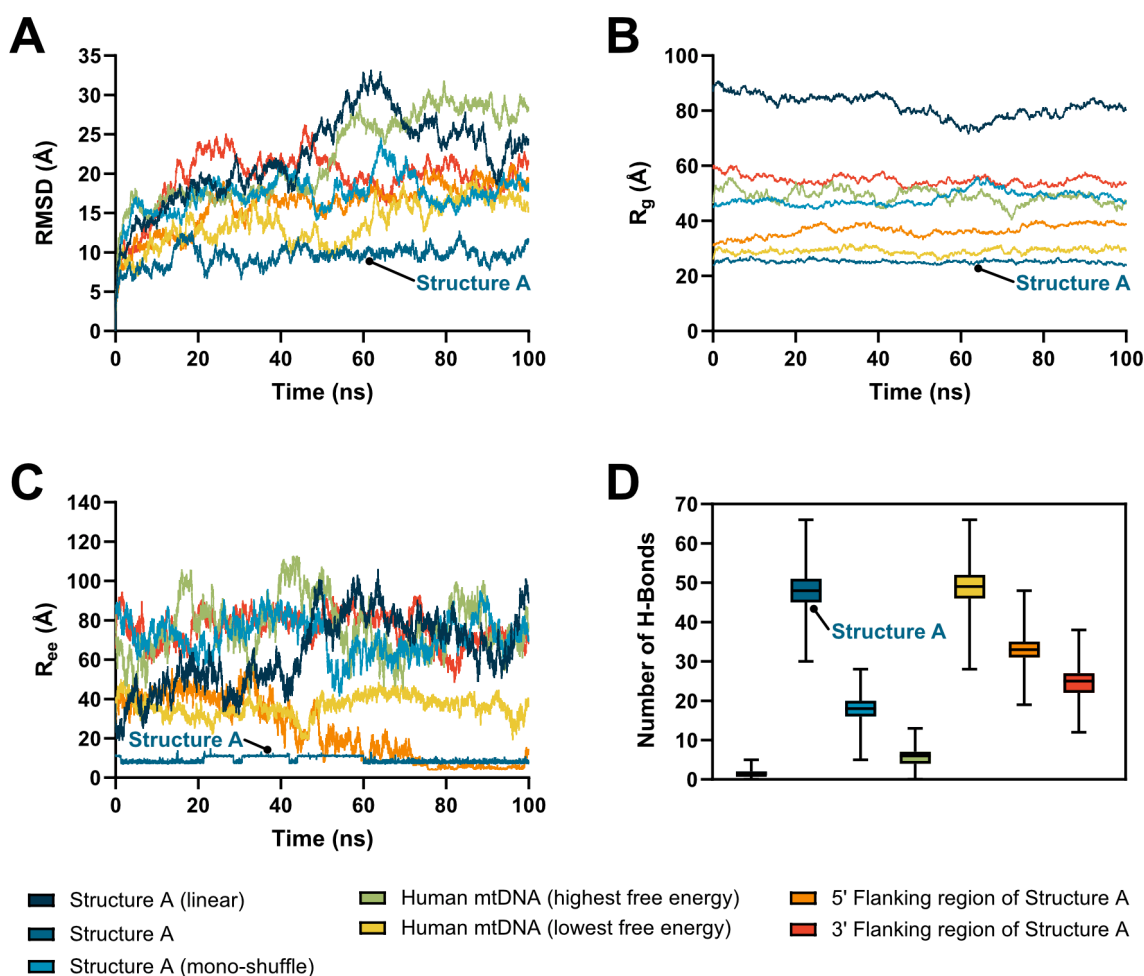


Figure 44. Analysis of the molecular dynamics simulation of the 3D Structure A conformation and controls. Graphs A, B, and C are represented as a function of time for the 100 ns of simulation. (A) RMSD values. All heavy atoms were used in the calculation of the mass-weighted RMSD. (B) Radius of gyration. (C) Distance between the centre of mass of the 3' and 5' ends of the DNA

conformations. (D) Box-and-whiskers plot of the number of hydrogen bonds established during the molecular dynamics simulation. The whiskers represent the minimum and maximum values, while the box extends from the 25th to 75th percentiles, with the line in the middle plotted at the median value.

Table 2. Summary of the results of the molecular dynamics simulations, excluding the first 10 ns. The results are represented as Mean (SD).

	Structure A	Controls					
		Structure A (linear)	Structure A (mono-shuffle)	Human mtDNA (highest free energy)	Human mtDNA (lowest free energy)	5' Flanking region of Structure A	3' Flanking region of Structure A
RMSD (Å)	10 (1)	23 (4)	18 (2)	23 (5)	14 (2)	17 (2)	20 (2)
R_{ee} (Å)	9 (2)	67 (16)	72 (10)	78 (14)	36 (5)	22 (16)	76 (8)
R_g (Å)	25.2 (0.6)	81 (4)	48 (2)	49 (2)	29 (1)	37 (2)	54 (1)
H-bonds	48 (4)	1.3 (0.8)	18 (3)	6 (2)	49 (4)	33 (4)	24 (3)

Even the 93-nt structure with lower MFE, considering the complete dataset of predicted 93-nt length structures in the mtDNA complete genome, presented mean RMSD values higher than those measured for Structure A. We also noticed that the nucleotide composition influences the non-B conformation formation, with the mono-shuffle prediction reaching a difference of approximately 8 Å higher than Structure A.

To visualise the conformational distribution that each structure acquired during the MD simulation, a two-dimensional space of the radius of gyration (R_g) and the end-to-end distance (R_{ee}) was created, where each point corresponds to a frame of the MD simulation (**Figure 45**). R_g and R_{ee} are complementary parameters that provide a rough estimate of the size and extension each structure acquires throughout the MD simulation (416). The density of the structures was represented by a heatmap, where hot areas (represented in red) are more densely populated than cold areas (represented in blue). Structures closer to the lower-left corner of the R_g - R_{ee} space are tightly wound and more compact. Contrarily, at the higher right corner, a structure is almost totally extended. Structures between these two regions are composed of several different sizes and shapes. As observed in **Figure 45**, Structure A has the most compact conformation, followed by the Human mtDNA structure with the lowest MFE. At the other end of the spectrum lies the linear Structure A with the highest R_g and R_{ee} . All the others are positioned near the centre of the R_g - R_{ee} space. Another interesting observation is that Structure A acquired two distinct conformations during the MD simulations, with an equal R_g but a higher R_{ee} .

These conformations result from the type of interaction that the two ends of Structure A establish. When the 5' nucleobase T1 establishes a Watson-Crick base pairing with the 3' nucleobase A93, this results in a higher R_{ee} . On the other hand, when the two ends perform a stacking van der Waals interaction, the R_{ee} is smaller since the centre of mass of these two residues is closer.

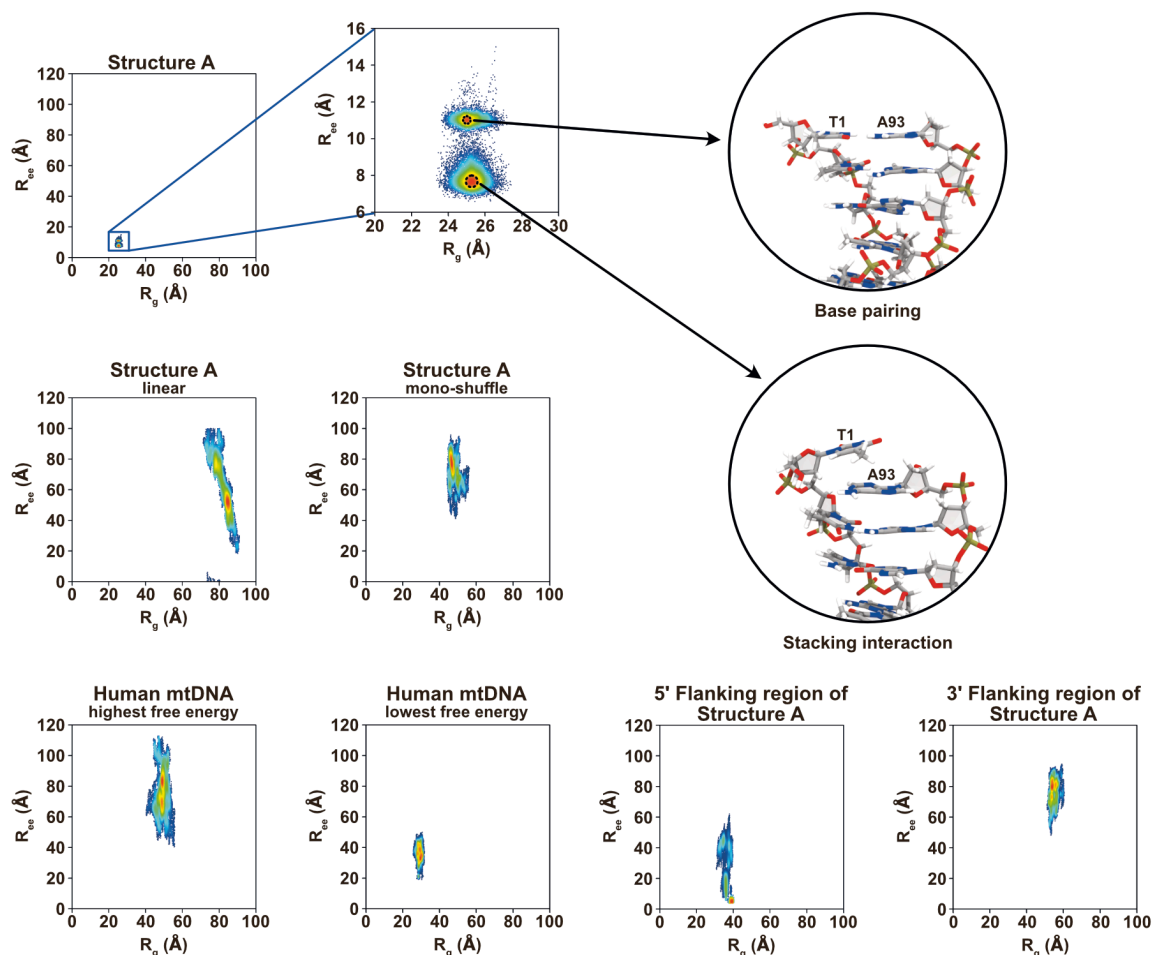


Figure 45. R_g and R_{ee} space of each structure is represented as a heatmap. Each frame of the MD simulations corresponds to one point in the graph, located by its (R_g , R_{ee}) value. The heat (i.e., colour) corresponds to the local density of points, with red being the most densely populated and blue the less densely populated. The R_g - R_{ee} space of Structure A is zoomed in to highlight the presence of two distinct populations. Representative structures of each of these populations are shown on the right. The main difference between the two populations is the nature of the interaction between T1 and A93 (i.e., 5' and 3' residues, respectively). The terminal residues of the structures belonging to the population with a higher R_{ee} establish Watson-Crick base pairing, whereas the ones with a lower R_{ee} perform van der Waals stacking interactions.

Regarding the number of H-bonds observed, Structure A presented a higher number ($n = 48$) than the other predictions, except for the structure with the lowest MFE, which presents only one additional H-bond ($n = 49$).

3.1.3.3. Structure A cluster analysis

We used a statistical method to ascertain the possible stability variation at region-specific segments corresponding to different structural elements. The criterion was the same for the global analysis of each structure previously described. The cluster analysis showed that the five different segments, obtained using the K-means method for Structure A, presented a mean RMSD of 10.3 Å (**Figure 46**). Considering the different clusters of Structure A, we obtained all segments with high stability, representing low MD RMSD variation between the initial and end states of all segments of the conformation. The other MD conformations simulations presented a mix between less stable segments and high stable structural elements. For instance, the Human mtDNA secondary structure with the lowest MFE and stability parameters, like Structure A, presented two segments with high RMSD variations (≥ 30 Å) and three segments with high stability parameters (lower RMSD variation, ≤ 10 Å). These difference in the stability of some regions of the mtDNA structure explains why local base-pairing can determine high stability patterns. These patterns are correlated with the stem-loop conformation in Structure A, while other conformations with multiple stem-loops present a variable degree of stability ($30 \text{ Å} \leq \text{RMSDs} \leq 10 \text{ Å}$). These high mean RMSD variations were also observed in the complete single-strand conformation, which we have used as a control.

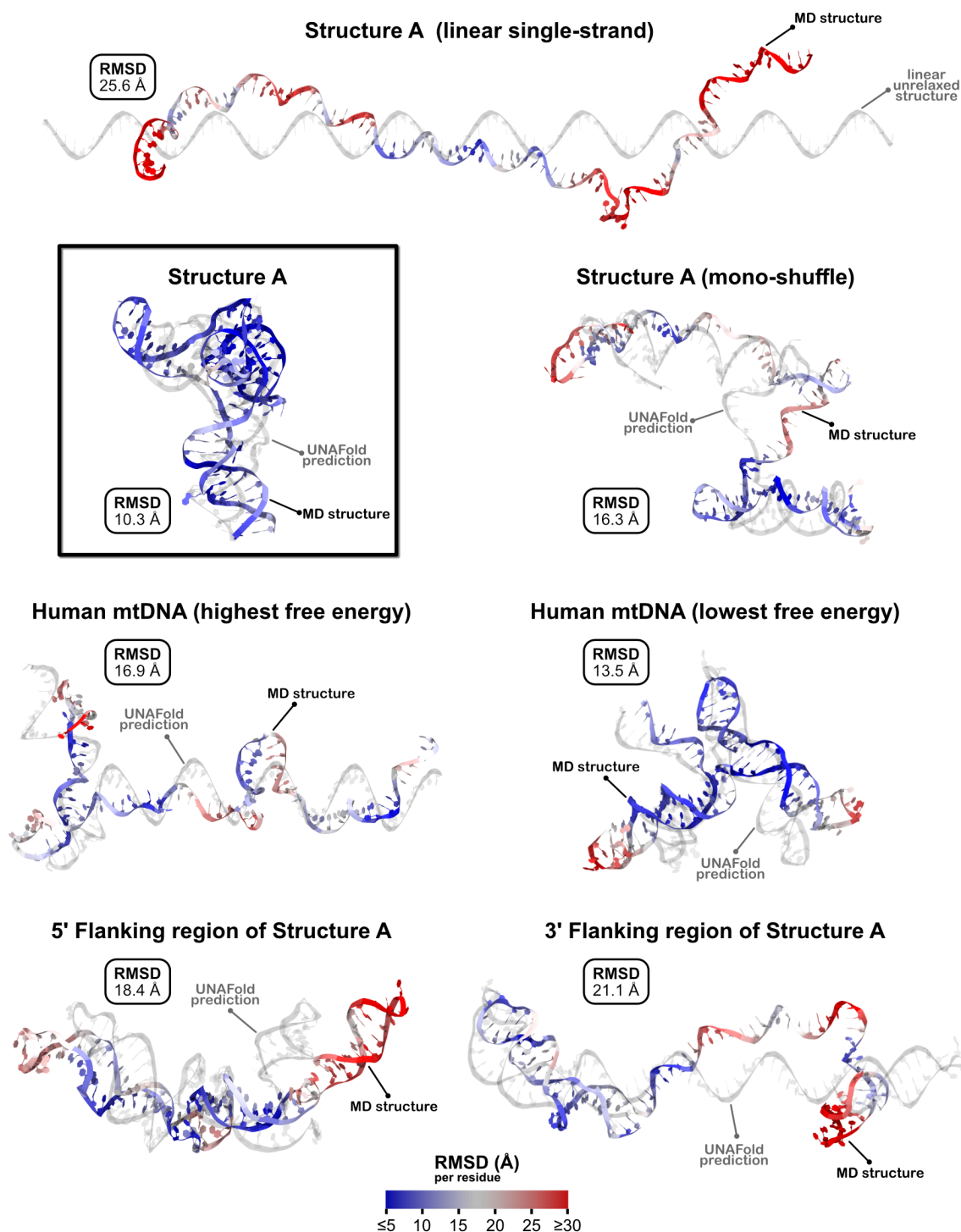


Figure 46. Superimposition of the representative structure of the major cluster obtained from the molecular dynamics simulation with the initial 3D prediction. The clustering was performed using the K-means method, and 5 clusters were obtained. The structures from the molecular dynamics simulations are coloured according to residue RMSD by comparison with the predicted structure.

3.1.3.4. Human, primate and mammalian conservation scores and RMSF variation

High conservation values imply that selective pressure is occurring to maintain specific residues or motifs unchangeable through evolutionary history (89). Structure A seems to have a relevant role in the modulation of mitochondria replication and during different stages of the diverse replication mechanisms (417). We assumed that the highest conserved segments of Structure A would show a low variation of mean RMSF values between the initial and end states of molecular simulation. In humans, all sites were polymorphic with a maximum of three variable nucleotides. The primate alignment has 82 sites out of 93, with no inter-species variation. The conservation scores for both the Human intra-species and primates alignment data were in accordance with previous results from Pereira et al. (89). Both alignments show all sites with at least two polymorphic sites (PIS = 0%). The PPI windows analysis (windows size = 1, step = 1) along Structure A revealed very high values in humans (PPI = 100%). The alignments of primates and mammals showed a higher difference between the minimum and maximum PPI values per position obtained. The primate lineage presented PPI values per position between 23% and 100%; and for the mammalian lineage, the PPI values per position were between 26.6% and 86%. The variation line tendency for the mean RMSF along the primate's alignment was remarkably like the PPI conservation scores (**Figure 47**), even though this correlation was only noticed for some regions of Structure A. A non-significant correlation was obtained in a global analysis (**Table 3, Figure 47**). In contrast, the highly stable 'ATGTA' stem region between positions 61 and 65 is invariant (100% conservation) in the primate alignment, which is in accordance with the conservation/RMSF negative correlation. In humans, the PPI conservation has high values (**Table 3**), and a negative correlation ($p < .01$) was observed when comparing the RMSF values and the PPI conservation scores, reflecting a pattern where the selective pressures are acting to preserve both the stability (low RMSD mean variation values) and the flexibility (low RMSF mean variation values). This pattern of negative correlation is observed at specific local regions, probably related to the different selective pressures occurring in different mammalian taxonomic groups. Considering the MFE analysis through different Chordata taxonomic groups, we have already determined that breakpoint regions in mtDNA have a low MFE (371) and are associated with non-B conformations. In this matter, the evolutionary analysis of Structure A using conservation scores also showed that global patterns of non-b confirmation occurrence are important to understand the mtDNA genome stability and local nucleotide changes inside each non-B conformation structure formed. Predicting the parameters associated with these nucleotide changes can help

build models that can be used for other regions of mtDNA associated with other diseases. Other studies have already used 3D DNA aptamers prediction to target cancer or detect toxic compounds (418,419).

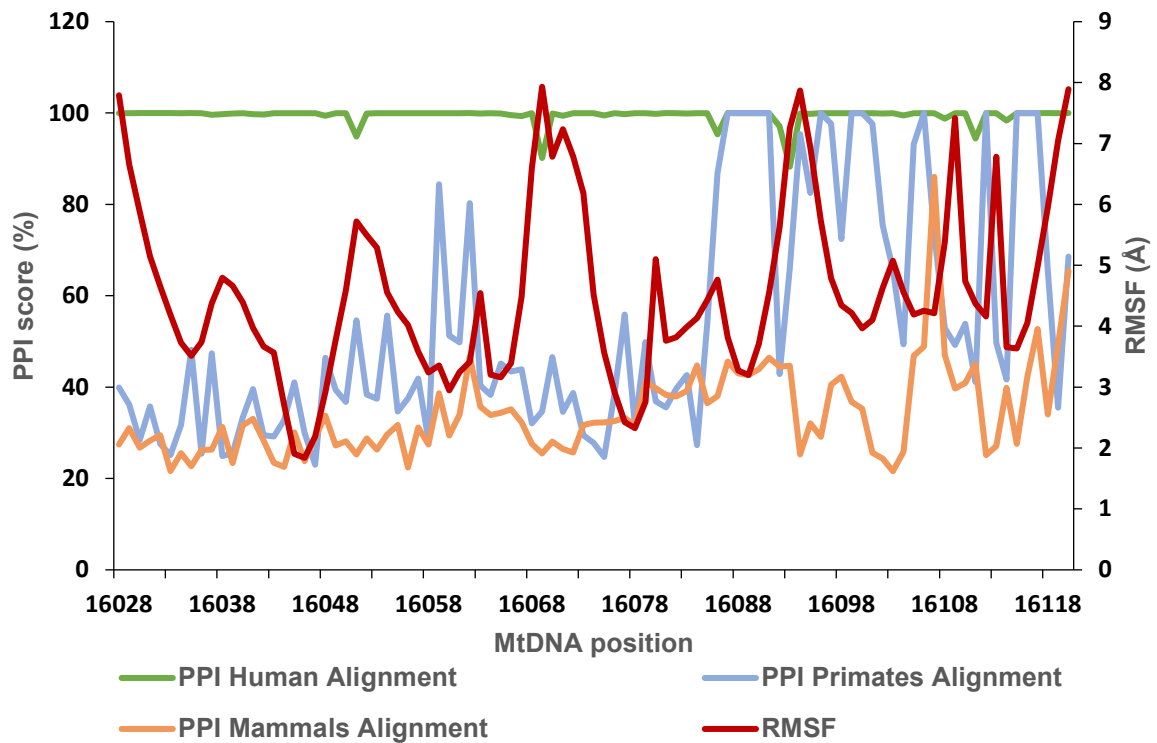


Figure 47. Analysis of the conservation scores of the 3D Structure A conformation considering the human intraspecific variation (Human Alignment) and the interspecific variation in Primates (Primates Alignment) and Mammals (Mammals Alignment). The horizontal axis shows the mtDNA nucleotide position, the primary vertical axis shows the conservation score percentage (%), and the secondary vertical axis shows the RMSFs values (Å). The red line represents the RMSF variation along the mtDNA structure A region.

Table 3. Summary of the Pearson's correlations, considering the conservation scores and the molecular dynamics RMSF of Structure A.

Variable 1	Variable 2	Pearson's r	p	Lower 95% CI	Upper 95% CI	
PPI Human Alig	PPI Primates Alig	0.019	0.855	-0.185	0.222	
PPI Human Alig	PPI Mammals Alig	-0.062	0.553	-0.263	0.143	
PPI Human Alig	RMSF Structure A	-0.31	**	0.003	-0.483	-0.113
PPI Primates Alig	PPI Mammals Alig	0.386	***	1.309e -4	0.198	0.547
PPI Primates Alig	RMSF Structure A	0.025	0.811	-0.18	0.228	
PPI Mammals Alig	RMSF Structure A	0.034	0.743	-0.17	0.236	

Table 3. Summary of the Pearson's correlations, considering the conservation scores and the molecular dynamics RMSF of Structure A.

Variable 1	Variable 2	Pearson's r	p	Lower 95% CI	Upper 95% CI
PPI Human Alig	PPI Primates Alig	0.019	0.855	-0.185	0.222
PPI Human Alig	PPI Mammals Alig	-0.062	0.553	-0.263	0.143
PPI Human Alig	RMSF Structure A	-0.31	**	0.003	-0.483
PPI Primates Alig	PPI Mammals Alig	0.386	***	1.309e -4	0.198
PPI Primates Alig	RMSF Structure A	0.025	0.811	-0.18	0.228
PPI Mammals Alig	RMSF Structure A	0.034	0.743	-0.17	0.236

* p < .05, ** p < .01, *** p < .001

3.1.3.5. Stack regions present higher conservation scores and low RMSF variation

The structures were analysed using as criteria for the conformation flexibility the mean of RMSF between the initial UNAFold predicted structure and the time of stabilisation. When analysing the structural element (e.g., stack, hairpin loop), the stem region defined by the stack base pairing showed the highest values of conservation for the human alignment data (**Figure 48**). This result is in line with previous Pereira et al. results that showed that structural stability highly depends on the number of complementary hydrogen-bonded nucleotides between paired strands (89). The RMSF's mean values between initial and end molecular dynamics conformations for the stack regions were very low (4.07 Å) compared to other structural elements (Table 4). The different structural elements of Structure A, like interior loops and hairpin loops, showed RMSF mean values of variation higher than those observed for stack and multi-loop regions (**Table 4, Figure 49**).

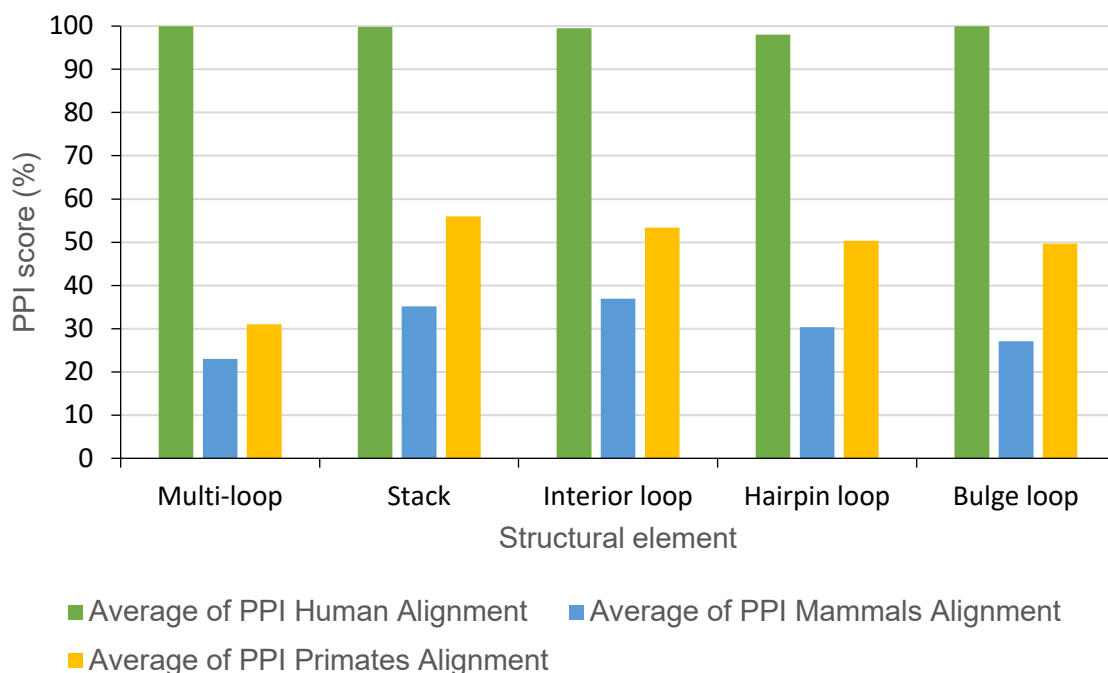


Figure 48. Analysis of the PPI conservation scores of the 3D Structure A conformation considering the structural element inside Structure A and the human intraspecific variation (Human Alignment), and the interspecific variation in Primates (Primates Alignment) and Mammals (Mammals Alignment). The horizontal axis shows the structural element, and the primary vertical axis shows the conservation score percentage (%).

Table 4. Descriptive statistics of Structure A structural elements analysis, considering the conservation scores and the molecular dynamics RMSF.

Structural Element	RMSF mean (Å)	PPI Human Alig. mean	PPI Primates Alig. mean	PPI Mammals Alig. mean	Sum of Disease-associated mutations	Sum of Somatic mutations
Bulge loop	6.80	99.95	49.66	27.08	0	0
Hairpin loop	6.28	98.01	50.31	30.34	2	5
Interior loop	4.64	99.49	53.38	36.96	0	0
Multi-loop	3.15	99.98	31.00	22.99	0	1
Stack	4.08	99.81	55.93	35.13	1	12

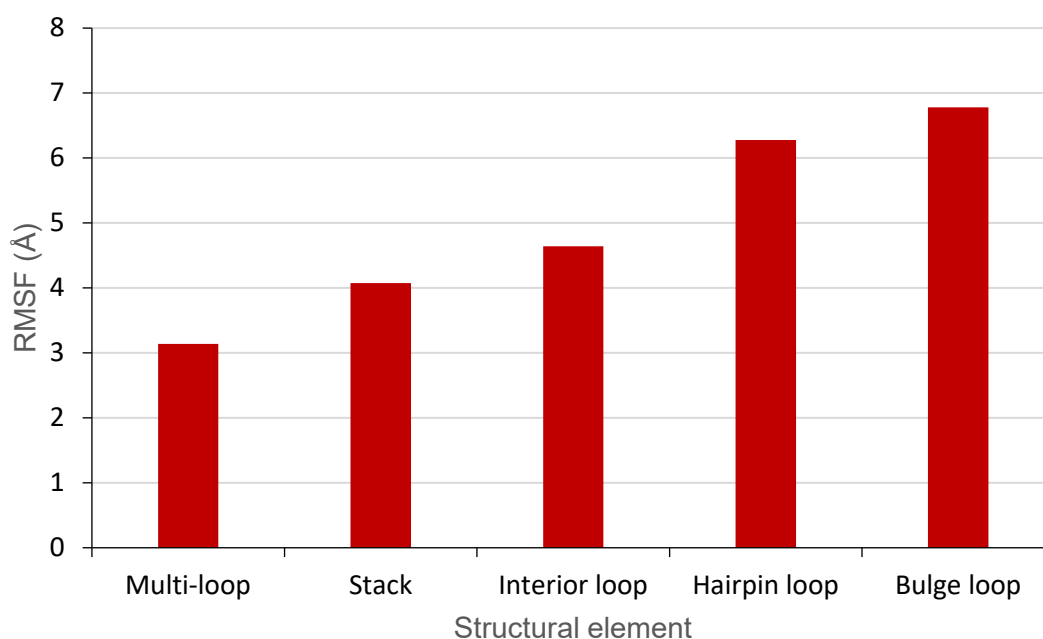


Figure 49. Analysis of the RMSFs values (Å) of the 3D Structure A conformation considering the structural element inside Structure A. The horizontal axis shows the structural element, and the vertical axis represents the RMSF values (Å).

3.1.3.6. Somatic and germline disease-associated variants within the Structure A elements

Although the Structure A region is non-coding mtDNA, some disease-associated variants have been described within this region (420–422). These variants were associated with V^{O2} max trainability, migraine headache and cyclic vomiting syndrome, and the syndrome of maternally inherited diabetes and deafness (MIDD). Our analysis of the correlation between each nucleotide conservation score of Structure A and human somatic variants retrieved a Pearson's statistical negative correlation ($p < .001$, **Table 5**). At the evolutionary level, both mammals and the sub-group primates retrieved no statistical correlation between PPI and RMSF mean values, suggesting that polymorphic states of Structure A region in humans, including deletions and insertions, can likely influence pathologic conditions. Further analysis of these diseases' pathways can give some clues to how Structure A influences the process, which to some extent, is putatively related to the mtDNA replication mechanism (95). These results have implications for the flexibility of mtDNA sites at position 16,068 at the hypervariable region I (HVR-I) and position 16,093, both located in hairpin loops, which show mean values of RMSFs (6.28 Å) between start and end conformations along the MD trajectory, higher than other structural

elements, and PPI (98.01%) conservation scores lower than other Structure A regions (Supplementary Material). The 16,068 (C16068T, $P = .02$) position was previously reported to occur in human low responders (LR) as defined by their maximal oxygen uptake response to a standardised aerobic exercise training program (421). We also observed the described Mitomap mtDNA somatic mutations at hairpin loops ($n = 5$), but surprisingly also at stack structural elements ($n = 12$). We predict that non-B DNA compensatory substitutions can occur at these sites, as shown in protein conformations (423,424), to maintain the structural stability of the molecular conformation.

Table 5. Summary of the Pearson's correlations, considering the conservation scores, the disease-associated mutations, and somatic mutations.

Pearson's Correlations			
Variable 1	Variable 2	Pearson's r	p
PPI Human Alig	Disease-associated	-0.052	0.622
PPI Human Alig	Somatic mutation	-0.553 ***	9.329e-9
Disease-associated	Somatic mutation	0.058	0.578

* $p < .05$, ** $p < .01$, *** $p < .001$

3.1.4. Conclusion

Our current understanding of all structural elements of the human genome is far from complete, but it has been finding a significant increase in the last few years. Non-canonical non-B structures are critical elements as they have been associated with genomic instability and, therefore, several pathological conditions. One of these structures, Structure A, was here in-depth analysed using a series of computational simulations, and the main results show that structure A is formed and remains stable during the simulations. The nucleotide positions with the highest conservation scores also show low RMSD and RMSF values, reflecting important selective pressures to maintain stability and flexibility. While waiting for future technical developments to aid undoubted proof of the formation of Structure A in an in vivo cellular environment, our computational analyses validate the possibility that these structural elements of the genome are key players in genomic instability.

This page intentionally left blank.

4. DNA aptamers

This page intentionally left blank.

4.1. Identification of new aptamers targeting cathepsin B – a protease overexpressed in prostate cancer

Cathepsin B (CatB) is a cysteine protease that has been found overexpressed in several different types of tumours, including prostate cancer. Hence, identifying novel DNA aptamers with high affinity for CatB may be promising for developing targeted therapies against those tumour cells. However, the methodology used to generate and isolate those aptamers (i.e., SELEX) is very laborious and can be expensive. For those reasons, *in silico* methods like molecular docking, molecular dynamics, and binding free energy calculations are starting to become attractive tools to help identify the best aptamers.

This work resulted from a strict collaboration with the group of Professor Lígia R. Rodrigues from the Centre of Biological Engineering – University of Minho, Braga, Portugal. André F. Pina performed the 3D predictions of all the DNA aptamers, their molecular dynamics simulations, docking with CatB, molecular dynamics of each aptamer-CatB complex, binding free energy calculations, and all the subsequent analyses. Regarding the article, André F. Pina wrote all the methodology, results, and discussion related to the *in silico* work mentioned above. The following content is almost a transcript of the published version, which was published in the Molecular Systems Design & Engineering journal.

Research Article

Identification of novel aptamers targeting
cathepsin B-overexpressing prostate cancer
cells

Ana Cláudia Pereira[‡], André F. Pina[‡], Diana Sousa[‡], Débora Ferreira, Cátia Santos-Pereira, Joana L. Rodrigues, Luís D.R. Melo, Goreti Sales, Sérgio F. Sousa, Lígia R. Rodrigues

Molecular Systems Design & Engineering, 7(6), 637–650

Impact Factor (2021): 4.920 [‡] The authors contributed equally to the work.

2022



10.1039/d2me00022a

Increased levels of cathepsin B (CatB), a cysteine protease, have been associated with different types of tumours, including prostate cancer. Hence, the identification of novel targeting ligands homing CatB may be a promising approach for the development of CatB-targeted therapies. In this work, a methodology called systematic evolution of ligands by EXponential enrichment (SELEX) was used to generate and isolate new aptamers

targeting CatB. After 8 rounds of selection, the pool was sequenced, and the 10 most prevalent sequences, along with their corresponding reverse complementary sequences, were further analyzed. In order to assess the binding affinity of the aptamers towards the CatB, bioinformatics tools were used. After generating the 3D structures for each aptamer, the HADDOCK software was used for the docking studies between them and CatB. Molecular dynamics simulations and free binding energy calculations by molecular mechanics-generalized born surface area (MM-GBSA) allowed the selection of a new aptamer with potential to target CatB. Cell binding assays against the PC-3 prostate cancer cell line confirmed the *in silico* predictions, further validating the newly discovered aptamer as a promising targeting ligand for CatB-overexpressing cancer cells.

This work was also presented at the following conference:

1. (Poster) *5th Meeting in Medicinal Biotechnology / 2nd Iberian Congress in Medicinal Biotechnology*, **2022**, Porto, Portugal (**Appendix**: page 267)

4.1.1. Introduction

Cathepsin B (CatB) has been gaining interest as a biomarker for tumour staging and malignancy assessment. It is a lysosomal hydrolase belonging to the cathepsin family, which contains three different groups: cysteine (B, C, H, F, K, L, O, S, V, W, X/Z), serine (G), and aspartate (D, E) cathepsins (425). Under physiological conditions, CatB is synthesized as an inactive precursor (pro-CatB) posteriorly activated in the acidic lysosomal lumen. Mature CatB has a heavy chain of 26 kDa and a light chain of 5 kDa, and unlike other proteases, it presents both endopeptidase, carboxypeptidase, and peptidyl-dipeptidase activities (425). This cysteine protease is considered a housekeeping protein involved in the proteolysis and, subsequently, recycling of cellular material. In normal cells, CatB is expressed in the lysosomes (426). However, in cancer cells, pro-CatB is first produced and secreted, yet remaining tied up to the cell surface via the Annexin II tetramer complex (425,427) and then released to the tumour microenvironment (425,426).

Several studies have demonstrated the potential role of CatB as a predictive cancer biomarker. For example, in a study where the serum levels of CatB in 260 patients were assessed (60 healthy controls, 80 with benign prostatic hypertrophy, and 120 with prostate cancer), increased levels of CatB were found in the serum of the patients with prostate cancer in comparison with healthy controls. Moreover, CatB levels correlated with disease extension as they were higher in patients with metastasis, as opposed to those without advanced disease (428). In another study, high CatB levels were associated with shorter survival of colorectal cancer patients (429). Similar conclusions have been reached in studies conducted for other types of cancer, such as bladder (430), lung (431), liver (432), breast (433), and mouth (434) cancers. Therefore, this protein appears as an excellent candidate for the development of targeting ligands and identifying targeting ligands homing CatB is a promising strategy for the development of novel targeted therapies.

Owing to their ability to bind to the target with high affinity and selectivity mediated by their tridimensional structure, aptamers have emerged as excellent biorecognition elements (435). They are oligonucleotide (RNA or single-stranded DNA) or peptide molecules, with lengths varying from 25 to 50 nucleotides (nt), that have been widely used to target proteins, organic and inorganic molecules, viruses, bacteria, and whole animal cells, being suitable for detection, imaging, diagnostic and therapeutic purposes (436–438). In contrast to antibodies, the gold standard for target recognition, aptamers are economical, highly reproducible, and easy to be chemically modified. Furthermore, due to their low

molecular weight, aptamers can easily infiltrate tissue and cells, typically having low toxicity and non-immunogenicity (439,440). Systematic Evolution of Ligands by EXponential enrichment, the so-called SELEX, is a technique that allows the isolation of aptamers from an initial oligonucleotide/peptide library by performing several iterative cycles, along which the pool increasingly enriches, thus leading to the identification of highly specific sequences for the chosen target (441).

In cancer research, with the improvement of the SELEX technique to reduce background, several aptamers have been identified as potential therapeutic tools against tumour cells (442,443). In fact, six aptamers (AX102, xPSM-A10 (A10), HB5, HeA2_3, MP7 and aptPD-L1) are in pre-clinical trials, and two (AS1411 and NOX-A12) are already in phase II clinical trials (444). AS1411 is an aptamer targeting nucleolin that can be used to either internalize the cell and inhibit DNA replication or as a drug delivery agent (443–445). NOX-A12 can serve as an antagonist, sensitizing CLL cells to cytotoxic drugs, improving treatment at lower doses (443–445).

In silico tools have been gaining increasing interest in the study of aptamers as therapeutic molecules because they can provide insights about the binding affinity between a given aptamer and the target of interest, at a fraction of the time, with reduced costs and satisfying accuracy (108). In fact, several studies in the field of oncology have applied computational methodologies, such as structure prediction, docking, and molecular dynamics, that were later corroborated by experimental studies, further validating the computational analysis. Indeed, several studies highlight the role and contribution of *in silico* tools to the development and optimization of new nucleic acid-based molecules (446–448). For instance, aptamers targeting the carcinoembryonic antigen (CEA) with ATG attached to the 30-end and GAC sequence attached to the 50-end were found to provide the highest score among the designed aptamers, with their high affinity being later confirmed experimentally by bilayer interferometry (449). In addition, two *in silico* studies focused on finding strong candidates against the epithelial adhesion molecule (EpCAM) also obtained aptamer sequences with high affinity that were later confirmed by experimental assays (446,448). In another study, Santini and colleagues used MD, MM-GBSA, and conformational analysis to isolate an aptamer sequence with potential for anti-cancer therapy by targeting the transmembrane glycoprotein mucin 1 (MUC1) (450). Another example was the use of the “rational truncation” technique to predict secondary structures of an aptamer targeting the prostate-specific membrane antigen (PSMA), where molecular docking approaches allowed the identification of the binding sites of the aptamers on the PSMA (451).

In this work, the SELEX technique was used to generate and isolate specific ssDNA aptamers as biorecognition elements homing CatB, aiming to contribute to the development of novel targeted therapies for CatB-overexpressing tumours. After aptamer selection, computational tools, including DNA-protein docking and free binding energy calculations, were employed to further explore the interaction of the candidates with CatB and select the best aptamer for further experimental validation.

4.1.2. Methodology

4.1.2.1. Reagents

Carboxylated magnetic beads MagSi-S COOH 3.0 (10 mg.mL⁻¹; 3.0 µm) were used from Gentaur, 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) and *N*-hydroxysuccinimide (NHS) were obtained from Sigma, and 2-(*N*-morpholino)ethanesulfonic acid (MES) buffer was purchased from Fisher Bioreagents. The binding buffer (BB) was prepared by supplementing phosphate buffered saline (PBS) 1×, pH 7.4, with 1 mM of CaCl₂·2H₂O (Panreac) and 0.5 mM of MgCl₂·6H₂O (VWR). The washing buffer (WB) was prepared by adding Tween 20 (OmniPur) at 0.05% (v/v) to BB and adjusting pH to 8. CatB was purchased from Abxbexa (abx069774). Primers and library were purified by high-performance liquid chromatography (HPLC) and purchased from Ella Biotech GmbH. Aptamer 1 was synthesized at Invitrogen and Aptamer 8RC at Metabion.

4.1.2.2. Preparation of CatB-conjugated magnetic beads

a. Magnetic beads surface activation

The magnetic beads (100 µL, 1–3×10⁹ beads.mL⁻¹) were first washed with 25 mM MES, pH 5. The suspension was then placed on a magnetic support for 5 min, and the washing solution was discarded. To activate the magnetic beads' carboxylic groups, solutions of EDC (50 µL, 50 mg.mL⁻¹) and NHS (50 µL, 50 mg.mL⁻¹) were prepared and then added to the washed beads. The suspension was incubated at room temperature under slow tilt rotation for 30 min. Next, the beads were washed twice with 100 µL of 25 mM MES pH 5.

b. CatB immobilization

Following activation, the beads were incubated with the target protein CatB (20 μL , 2 $\mu\text{g}\cdot\mu\text{L}^{-1}$) overnight at room temperature under gentle stirring conditions. Following incubation, the suspension was placed on a magnetic support for 5 min to remove any unbound protein. To block unreacted carboxylic groups of the beads' surface, 100 μL of 50 mM ethanolamine in PBS 1 $\times$ was added to the functionalized beads and further incubated at room temperature under gentle stirring for 1 h. After incubation, the excess of ethanolamine solution was removed, and beads were washed three times with 100 μL of WB. The CatB-conjugated beads (positive matrix) were resuspended in PBS 1 $\times$ and stored for up to a week at 4 °C. The same procedure was performed without CatB to obtain the negative matrix used in the selection rounds.

4.1.2.3. SELEX procedure

To initiate the SELEX process, an ssDNA library pool (500 pmol) comprised of 50 nt in the random region and flanked by primer binding sites on both sides (ATCCAGAGTGACGCAGCA-50 nt-AATGTCGTTGGTGGCCC) was heated to 95 °C for 5 min, cooled on ice for 15 min, and then left at room temperature for another 15 min. Next, the pool was suspended in 25 μL of the negative matrix (negative selection), and BB was added to a final volume of 100 μL . This mixture was then incubated at room temperature with gentle shaking for 45 min. Afterwards, the suspension was placed in the magnetic support, and the recovered supernatant was incubated with the positive matrix (positive selection) under the same conditions for 45 min. Following incubation, the unbound ssDNA was removed. The beads were resuspended in ultrapure H_2O , heated at 95 °C for 3 min to detach the ssDNA bound to CatB, and the supernatant was then recovered, quantified, and amplified by PCR (see the following section). To increase the stringency of the process, decreasing concentrations of CatB were used as follows: 8 μg were used for the 1st round, 6 μg for the 2nd and 3rd rounds, 4 μg for the 4th round, 2 μg for the 5th and 6th rounds, and 1 μg was used in the last two rounds. Before every round, a negative selection round was performed to exclude any ssDNA unspecific binding to the magnetic beads.

4.1.2.4. Amplification and ssDNA pool purification

The PCR amplification was performed using primers complementary to the primer binding sites of the library randomized region using the following primers: Fw, 5'-ATCCAGAGTGACGCAGCA-3' and Rv, 5'-GGGCCACCAACGACATT-3'. In each positive round, 10 ng of ssDNA pool was used in a Master Mix containing 0.4 μ M of each primer, 12.5 μ L of HIFI DNA polymerase 2 $\times$ Master Mix (VWR) and RNase free H₂O to a final volume of 25 μ L. After 5 min at 95 °C, 25 cycles (one less cycle in each of the subsequent rounds) consisting of 30 s at 95 °C for denaturation, 30 s at 72 °C for annealing, and 20 s at 72 °C for elongation were performed, with a final step of 5 min at 72 °C. The PCR products (85 nt) were analyzed on a 2.5% (w/v) agarose gel to check the size of the dsDNA products. The PCR product was purified by MinElute PCR purification kit (Qiagen) and denatured at 95 °C for 5 min for the next round of selection.

4.1.2.5. Next-generation sequencing and data analysis

The resulting ssDNA pool after the 8th round was sent for sequencing at the Stab Vida laboratory, where it was prepared by performing an overlap PCR with primers complementary to constant regions of the randomized oligonucleotide library with an added overhang that includes a platform-specific sequence. The sequencing reads were then filtered and demultiplexed. For this analysis, constant primer binding regions were removed, and sequences longer than 55 nt and shorter than 45 nt were discarded. Raw next-generation sequencing (NGS) data was analyzed using the alignment algorithm embedded in Geneious 9.1.4 software (Biomatters Ltd.; <https://www.geneious.com/>). The 10 most frequent oligonucleotide sequences, as well as their reverse complementary ones, were further evaluated. The multiple alignment of the 20 selected sequences was performed on Geneious 9.1.4 using a MAFFT alignment with the FFT-NS-2 algorithm. The tree was generated using the Geneious tree builder employing the Tamura–Nei genetic distance model and the neighbour-joining method. The tree was built on Geneious 9.1.4, using a bootstrapping of 1000.

4.1.2.6. Generation of three-dimensional DNA aptamers models

The generation of three-dimensional (3D) DNA aptamer models, starting from its nucleotide sequence, consisted of three sequential steps, as summarized in **Figure 50**.

The overall idea of the present workflow was based on the work of ref. (402) with some alterations as described below (402).

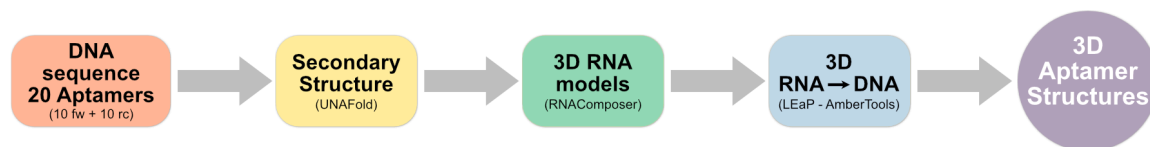


Figure 50. Summary of the workflow used to generate the DNA aptamer models, from their nucleotide sequence to their 3D structure. fw, forward; rc, reverse complement; UNAFold, unified nucleic acid folding, and hybridization package.

The first step consisted of predicting the secondary structure of the DNA aptamers, starting from their nucleotide sequence. The DNA sequence of the 20 aptamers (the 10 top sequences and respective complementary sequences) under study was used as input for the Mfold web server (452), currently part of the UNAFold web server (379). This web server predicts the secondary structure of single-stranded nucleic acids using free energy minimization algorithms. The application “DNA Folding Form” of the Mfold web server was used. The DNA sequences were defined as linear, with a folding temperature of 25 °C. The default ionic conditions of 1 M of Na⁺, and 0 M of Mg²⁺, with the correction for oligomers, were maintained. Only foldings within 5% from the minimum free energy were computed. A maximum number of 50 foldings were computed, with no limit regarding the maximum distance between paired bases. The resultant structure for each nucleotide sequence with the lowest free energy was then used in the subsequent steps. Secondary structures in the dot-bracket notation (Vienna format) obtained from the Mfold analysis were then used as input for the RNA Composer web server to generate the starting 3D models (400,401). Due to the lack of available tools for the prediction of 3D single-stranded DNA (ssDNA) structures (402), an equivalent 3D RNA structure was generated instead. The resulting tridimensional structures of RNA were then downloaded in the protein data bank (PDB) file format. In order to convert them back to DNA, an in-house bash script was used to remove the hydroxyl group at the C2' position of the ribose ring and the hydrogen atom at the C5 position of uracil. Next, the LEaP program, part of the AmberTools software package (110), was used to add all the missing atoms, including the methyl group needed to convert uracil into thymine. The resulting structures represented 3D DNA molecules suitable for the next steps of the computational pipeline used in this work.

4.1.2.7. Molecular dynamics simulations of the aptamers

Molecular dynamics (MD) simulations were made for the 20 modelled DNA aptamers. In order to prepare them for MD simulations, the aptamers were first solvated with TIP3P (286) water molecules in a rectangular box so that each atom of the aptamer is distanced at least 12 Å from the edge of the periodic box. Then, a physiological concentration of NaCl (150 mM) was introduced in the system by the Leap package of the AmberTools18 software (110). The number of Na⁺ and Cl⁻ ions added was determined using the method screening layer tally by container average potential (SLTCAP) (405). The ions were added in two steps. Firstly, enough Na⁺ ions were added to neutralize the DNA, and then both Na⁺ and Cl⁻ were randomly added to reach the desired concentration. The energy minimizations and MD simulations were performed with bsc1 (406), a force field developed for the atomistic simulation of nucleic acids. The minimization of the system was performed in four steps using the sander program of the AMBER18 software package: (1) water molecules; (2) hydrogen atoms; (3) all non-backbone atoms; and (4) all atoms. Then, the system was equilibrated in two stages using the NVT ensemble. Firstly, the system was gradually heated from 0 to 298.15 K, over 20 ps, while applying weak restraints on the DNA aptamer. Then, the restraints were removed, and the system was maintained at this temperature for another 80 ps.

The production phase of the MD simulations was run for 50 000 000 steps, with an integration time of 2.0 fs, totalling 100 ns. The SHAKE (139) algorithm was used to constraint all bonds involving hydrogen atoms. The system was maintained at constant temperature (298.15 K) and pressure (1.0 bar) using the Langevin thermostat (407) and Berendsen barostat (146), respectively. A cutoff of 10 Å was applied to the non-bonded interactions. The equilibration and production phases were conducted using the PMEMD software from the AMBER18 package.

4.1.2.8. Aptamer-CatB Docking

Five representative structures from the MD simulations were chosen for each of the 20 aptamers to perform molecular docking with the CatB protein. The choice of such structures was performed through cluster analysis. Using the K-means clustering method (453,454), five clusters were obtained for each aptamer, having the root mean square deviation (RMSD) of all non-hydrogen atoms as a measure of similarity. The representative structure of each cluster was then chosen to dock with CatB.

The HADDOCK 2.4 web server (455) was used to predict the complexes between CatB and the 20 aptamers under study. Since no experimental evidence regarding the potential binding sites is available, no active or passive residues were defined. Instead, ambiguous interaction restraints (AIRs) were randomly defined from residues holding >20% relative accessibility by selecting the random patches option. For each docking run, a set of different residues from CatB and the aptamer were randomly selected. The sampling was 10 000 structures for the it0 stage, 400 for the it1, and 400 for the itw. The fraction of common contacts (FCC) clustering metric, which calculates the fraction of inter-protein residue pairs common between complexes (456), was used with a cutoff of 0.6 and a minimum cluster size of 4. The ranking of the different complexes for each aptamer was established using the HADDOCK score (HS), a linear weighted sum of energetic terms and buried surface area ($HS = 1E_{vdw} + 0.2E_{elec} + 1E_{desol} + 0.1E_{AIR}$). Each term corresponds to van der Waals, Coulomb electrostatics, desolvation, and restraint energies, respectively (185,186). For each aptamer, the top-ranked docked complex was selected for further analysis through MD simulations.

4.1.2.9. Molecular dynamics simulations of the docked CatB–aptamer complexes

The MD simulations of the top-ranked docked CatB-aptamer complexes were run as previously mentioned for the initial simulation of the 20 aptamers (**section 4.1.2.7**). All the parameters were maintained except for the additional use of the ff14SB force field (129), employed in the simulation of proteins.

4.1.2.10. MM-GBSA method for binding free energies calculations

The binding free energies (ΔG_{bind}) for each CatB-aptamer complex were predicted using the molecular mechanics–generalized born surface area (MM-GBSA) method with the MM-PBSA.py script (457) from AMBER 18 (110). This method exhibits a strike balance between computational cost and reliability (458) and has been broadly used to predict the binding affinity of several biomolecular complexes (e.g., protein–protein, protein–DNA, protein–drug, and DNA–drug), as well as to evaluate and rescore docking solutions (459–462). For this analysis, 500 structures retrieved from the last 50 ns of each complex MD simulation were used.

The difference between the total free energy variation when CatB and the aptamer are in the bounded or non-bounded states is used to calculate the total binding free energy (**Equation 62**).

$$\Delta G_{bind} = \Delta G_{complex} - (\Delta G_{CatB} + \Delta G_{Apt}) \quad (62)$$

If $\Delta G_{complex}$ is lower than the sum of the ΔG of the separated protein and aptamer, the ΔG_{bind} is negative, meaning that the association of CatB with the aptamer under study is thermodynamically favourable. The ΔG_{bind} can be further described by the sum of different energy terms (**Equation 63**):

$$\Delta G = \Delta E_{gas} + \Delta G_{sol} - T\Delta S \quad (63)$$

$$\Delta E_{gas} = \Delta E_{int} + \Delta E_{ELE} + \Delta E_{VDW} \quad (64)$$

$$\Delta G_{sol} = \Delta G_{GB} + \Delta G_{Surf} \quad (65)$$

The gas-phase interaction energy (ΔE_{gas}) between CatB and the aptamer is the sum of electrostatic (ΔE_{ELE}) and van der Waals (ΔE_{VDW}) interaction energies. Since CatB, the aptamer, and the complex CatB-aptamer were extracted from the same trajectory, the internal energy change (ΔE_{int}) is annulled (**Equation 64**). Polar and non-polar solvation free energy terms contribute to the solvation free energy (ΔG_{sol}) (**Equation 65**). The polar solvation free energy (ΔG_{GB}) is determined by a modified generalized born (GB) method (463). The non-polar contribution to the solvation free energy (ΔG_{Surf}) is based on the solvent accessible surface area (SASA) calculation by the linear combinations of pairwise overlaps (LCPO) method (190). The entropy term ($T\Delta S$) was not considered in our calculations as it tends to introduce low accurate approximations (457).

4.1.2.11. Aptamers' characterization

The aptamer 1 (Apt01) and its complementary sequence (Apt08rc) were synthesized with a fluorescent report at 5' end (Apt1: 5'-FAM-ATCCAGAGTGACGCAGCATGGGGGGCAAGCAATTGGTGTGTGGTGTAGCATCCCCGTGTACCGCGCGGGAATGTCGTTGGTGGCCC-3'; Apt08rc: 5'-FAM-GGGCCACCAACGACATTGGGCAGGCAACCTCCCACTATTGAAGACACTACCACCACCACTGCACACATGCTGCGTCACTCTGGAT-3'). Then, a calibration curve for each aptamer was performed using the following range of concentrations (0-100 nM) in PBS 1×. The standard solutions were prepared, then heated to 95 °C for 5 min, followed by 10 min on

ice and an additional 10 min at room temperature before transferring solutions to a 96-well plate. Fluorescence excitation was done at 490 nm, and emission was read at 525 nm in a microplate reader (Cytation 3, BioTek).

To determine the aptamers' dissociation constant (K_d), calibration curves with a fixed concentration of aptamer and increasing concentrations of CatB (0-100 nM) were conducted, and fluorescent signal assessment was performed at 525 nm. The software Origin (Pro) 2021 (OriginLab Corporation, Northampton, MA, USA) was used to fit the data to a four-parameter logistic curve using the Hill 1 Model with **Equation 66**.

$$y = \text{start} + \frac{(\text{end} - \text{start}) \times x^n}{k^n + x^n} \quad (66)$$

4.1.2.12. Cell-Binding assays

Prostate cancer (PC-3) cells (CRL-1435; ATCC) were cultured in Dulbecco's Modified Eagle Medium (DMEM) (Biochrom) supplemented with 10% fetal bovine serum (FBS) (Biochrom) and 1% penicillin–streptomycin (Biochrom) on tissue culture treated flasks (Sarstedt) and propagated at 37 °C, 5% CO₂ and 95% relative humidity. The cells were washed with PBS 1× pH 7.4 and were detached using Trypsin/EDTA solution (Biochrom). For the cell-binding assays, cells were first seeded on 96-well plates at a cell density of 1×10⁴ cells per well overnight at 37 °C. Afterwards, the medium was removed, and the cells were washed with 100 µL of PBS 1×. Next, standard concentrations of aptamer (0, 10, 50, 100, 250, and 500 nM) were incubated with PC-3 cells for 1 h at room temperature. All assays were performed in duplicates. After incubation, unbound aptamers were removed, and cells were washed with PBS 1×. Then, cells were detached and transferred to a black 96-well plate. Fluorescence was measured at 525 nm in a microplate reader (Cytation 3, BioTek).

4.1.2.13. Statistical analysis

The statistical analysis for the cell-binding assays was performed using GraphPad Prism 5.03 (GraphPad Software, Inc.). Statistical significance was evaluated using two-way ANOVA with Tukey's multiple comparison test, with a 95% significance level.

4.1.3. Results and Discussion

4.1.3.1. Selection of aptamers against CatB aided by NGS sequencing

The SELEX technique was chosen to identify DNA aptamers with high affinity towards CatB. For this purpose, a library of 50 random sequences flanked by constant regions was used. After 8 positive selective rounds intercalated with 8 counter-selective cycles (negative selection), the pool of aptamers was sequenced by NGS. Out of 4 823 876 sequences, the 10 with the highest number of copies were retrieved. As we did not perform strand separation after the SELEX procedure, the reverse complementary sequences of the top 10 aptamers were also considered (

Table 6).

Table 6. Top 10 sequences and their complementary sequences (bold) with the highest number of copies retrieved from the total number of sequences obtained by next-generation sequencing (NGS).

Aptamer ID	5' - Variable Sequence -3'	Number of copies
Apt01	TGGGGGGCAAGCAATTGGTGTGTGGTGTAGCATCCCCCTGTACCGCGCGGG	46
Apt01rc	CCCGCGCGGTACAGGGGGATGCTACACCACACACCAATTGCTTGCCCCCA	
Apt02	TGTCGGGGGTGGTGTGTTGTGCAAAATGTTGGGATGTGTCCGGGGTGGGC	45
Apt02rc	GCCCCCGGACACATCCCAACATTTTGCACAACACACCACCCCGACA	
Apt03	TACGGGGGAGGAGTAAATGGGGTGAGGTAATGTGGTGTACAGGTGGCGGG	40
Apt03rc	CCCGCCACCTGTACACCACATTACCTCACCCCATTTACTCCTCCCCCGTA	
Apt04	TGTGAAGTGCAGGCACAATCGGTGGTGTGGCGATGTTCCAGTGCACGCC	29
Apt04rc	GGGCGTGCACCTGGAACATCGCCACACCACCGATTGTGCCTGCACCTCACA	
Apt05	TGGTGGCAAGCAGTGTGTTACATATTACTTGACCTACTGCACCTTTGCC	29
Apt05rc	GGGCAAAGGTGCAAGTAGGTCAAGTAATATGTAACACACTGCTTGCCACCA	
Apt06	TGCCGGGGAGGAGTTCGTGCGGAAGTCGTTCCATAACACCGCTAAGGCC	28
Apt06rc	GGGCCTTAGCGGTGTTATGGAACGACTTCCGCACGAACTCCTCCCCGGCA	
Apt07	TGGTGAGGGTAAGGTGAGGATTTGTGGTACAGGGTTGGTATCGCAGGCG	27
Apt07rc	CGCCTGCGATACCAACCCTGTACCACAAATCCTCACCTTACCCTCACCA	
Apt08	TGTGTGCAGTGGTGGTGGTAGTGTCTTCAATAGTGGGAGGTTGCCTGCCC	26
Apt08rc	GGGCAGGCAACCTCCCACTATTGAAGACACTACCACCACCACTGCACACA	
Apt09	TGGTGGGGAGGAAGTATGGTTGTCTTTAGGTAAAGTGTGTGTACGCGCCC	26
Apt09rc	GGGCGCGTACACACACTTAACCTAAAGACAACCATACTTCCTCCCCACCA	
Apt10	TGTGGGCAGATTAGTGGTTATCCTTGCTGTTGGTTGGTATGCGCCC	26
Apt10rc	GGGCGCATACCAACCAACAGCAAGGATAACCACTAATCTGCCCACA	

The 20 sequences were further aligned to assess homology within the variable region regarding sequence alignment and phylogenetic relationship (**Figure 51**). A sequence similarity above 50% was verified between Apt04rc–Apt01rc/Apt02rc, Apt03–Apt07, Apt10–Apt09, Apt03rc–Apt07rc, Apt10rc–Apt08rc/ Apt09rc, and Apt09rc–Apt08rc. On the opposite, three combinations of aptamers showed sequence similarity below 10%, Apt02rc–Apt03/Apt07/Apt06 (**Figure 51A**). Furthermore, the phylogenetic tree shows that the aptamers that were chosen below for experimental validation (Apt01 and Apt08rc) display distinct paths. Four aptamers, Apt02rc, Apt05rc, Apt07rc, and Apt07, presented the most dissimilar sequences (**Figure 51B**).

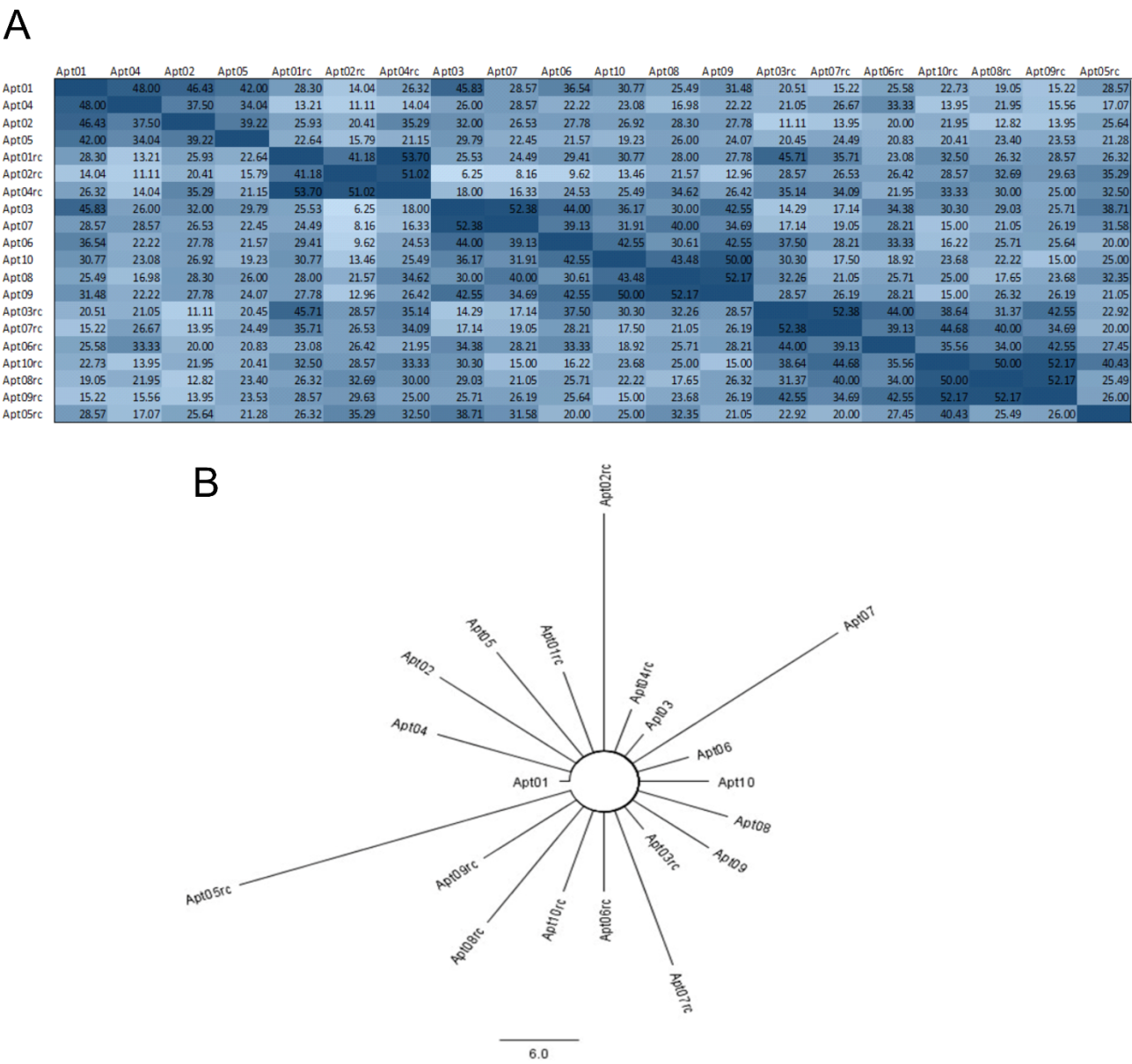


Figure 51. Primary structure and phylogenetic tree analysis. The random region (50 nt) of the ten top-ranked oligonucleotide sequences, along with their respective reverse complementary sequences obtained after NGS, were evaluated by performing (A) multiple sequence alignments on

Geneious 9.1.4 using a MAFFT alignment with the FFT-NS-2 algorithm (% of similarity), and (B) phylogenetic analysis generated using the Geneious tree builder employing the Tamura–Nei genetic distance model and the neighbour-joining method.

4.1.3.2. Analysis of aptamers MD simulation

A multi-step computational pipeline was employed to evaluate the different aptamers obtained by SELEX regarding their binding affinity to CatB (**Figure 52**). The 3D structures of the 20 selected aptamers were first predicted to allow the second stage, which was the CatB-aptamer docking. Then, the CatB-aptamer complexes were further refined through MD simulations and MM-GBSA calculations to obtain a final ranking based on the Haddock score and ΔG_{bind} , which allowed the selection of an aptamer for experimental validation.

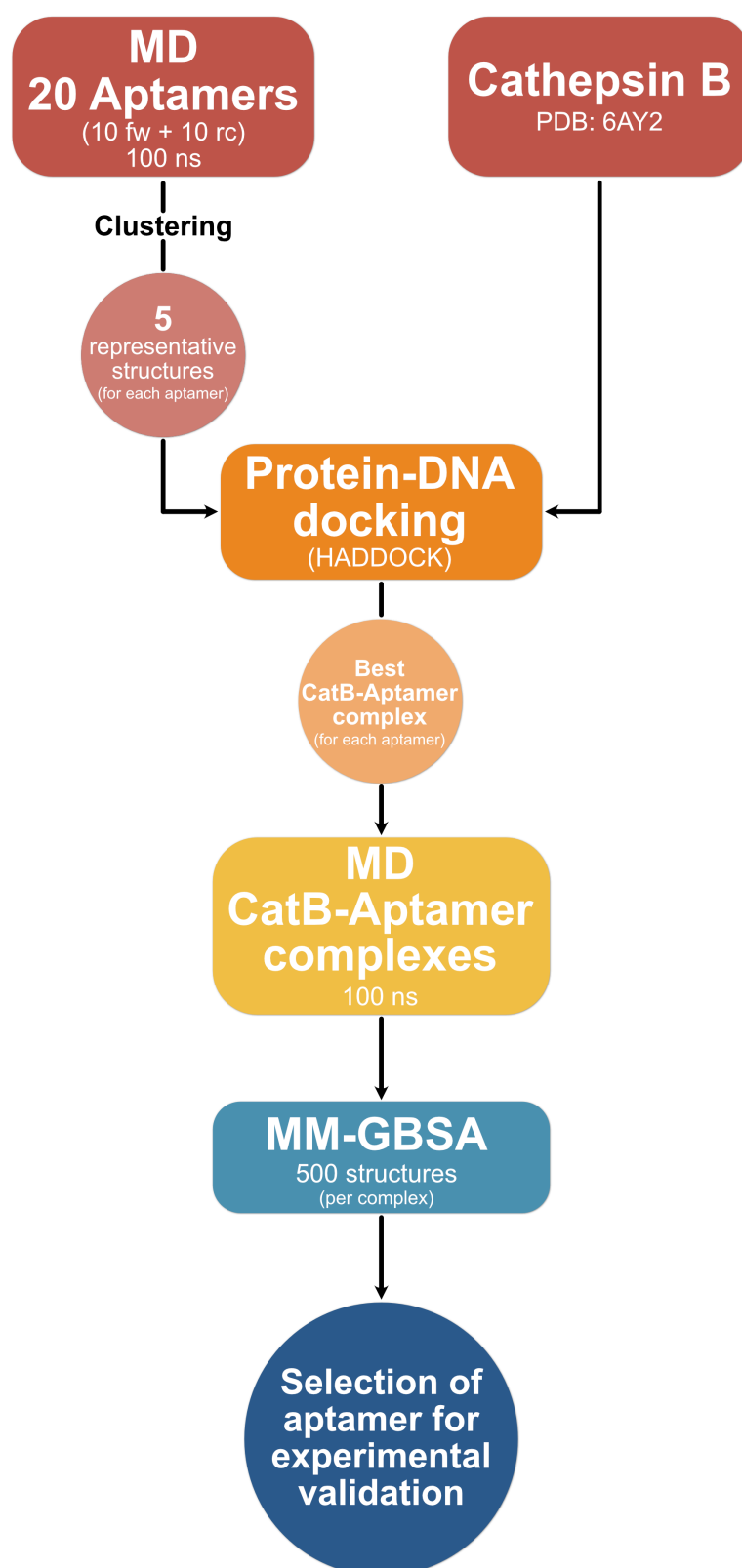


Figure 52. Flowchart of the workflow applied in this work starting from the generated 3D models of the DNA aptamers. fw, forward; rc, reverse complement; MD, molecular dynamics; HADDOCK, high ambiguity driven protein–protein DOCKing; CatB, cathepsin B; MM-GBSA, molecular mechanics-generalized born surface area.

All the generated 3D aptamer models were minimized, equilibrated, and further submitted to 100 ns MD simulations as a way to sample a high number of conformations. Posteriorly, the structures were grouped into 5 clusters by the K-means clustering method, according to their RMSD of all non-hydrogen atoms, used as a measure of similarity. The representative structure of each cluster was then chosen for molecular docking with CatB. This procedure was performed for each of the 20 aptamers. All the aptamers had high RMSD values throughout the simulations, never lower than 10 Å and reaching values as high as 30 Å (**Supplementary Figures S1-S4**). From all the structures evaluated, only the aptamers Apt01, Apt04, Apt06, Apt07, and Apt10 presented a relatively stable RMSD of all non-hydrogen atoms. This behaviour is expected for single-stranded DNA molecules, which have higher conformational flexibility than double-stranded DNA, especially for longer polynucleotide chains, such as the aptamers under study. Indeed, simulations with other ssDNA aptamers exhibited similar increased RMSD values (464), agreeing with our data.

4.1.3.3. CatB-Aptamers molecular docking

In order to predict the complexes between CatB and the 20 aptamers under study, the HADDOCK 2.4 software was used. Since no experimental data concerning the binding pocket was available, the *ab initio* mode, which considers all the solvent-accessible residues harbouring more than 20% relative accessibility, was chosen. These calculations were carried out using five different structures for each aptamer, generating numerous CatB-DNA complexes for each aptamer, from which the top complex obtained after the final refinement in water (itw stage) was selected for further analysis. All the complexes obtained had a negative Haddock score ranging from -155.6 to -85.5 a. u. and a buried surface area varying from 1314.1 to 3215.5 Å² (**Table 7**, 2nd and 3rd columns). Considering the Haddock scoring function, complexes formed by CatB and aptamers Apt03, Apt05, Apt06, Apt07, Apt04rc, Apt08rc, and Apt09rc are those standing out for having a score below -110 a. u.

Table 7. Haddock results and binding free energy calculations. The energies were calculated using 500 structures of the last 50 ns of the MD simulations of the CatB-Apt docked complexes using the MM-GBSA method. The polar contribution to the solvation free energy (ΔG_{GB}) and the electrostatic energy (ΔE_{ELE}) are accounted in the polar ΔG , while the non-polar ΔG refers to the non-polar contribution to the solvation free energy (ΔG_{surf}) plus the van der Waals energy (ΔE_{VDW}). Haddock score is provided in arbitrary units (a. u.), buried surface area in Å², and the energy values in

kcal.mol⁻¹ ± standard deviation. The aptamer chosen for experimental validation is highlighted in bold.

Aptamer	Haddock score (a.u.)	Buried surface area (Å ²)	ΔG_{bind}^{Total} (kcal.mol ⁻¹)	ΔG Non-polar (kcal.mol ⁻¹)	ΔG Polar (kcal.mol ⁻¹)
Apt01	-105.1	1764.5	-19.8 ± 0.6	-111.1 ± 0.6	91.3 ± 13
Apt02	-107.2	1869.5	-26.6 ± 0.5	-65.7 ± 0.4	39.1 ± 24
Apt03	-124.6	3140.4	-37.6 ± 0.5	-80.2 ± 0.6	42.6 ± 25
Apt04	-90.3	1834.9	-71.5 ± 0.5	-89.2 ± 0.4	17.7 ± 17
Apt05	-117.3	1929.0	-45.6 ± 0.3	-49.3 ± 0.2	3.6 ± 24
Apt06	-155.6	3043.5	-51.7 ± 0.5	-106.3 ± 0.8	54.6 ± 19
Apt07	-116.0	1963.4	-46.3 ± 0.5	-136.6 ± 0.5	90.3 ± 20
Apt08	-109.2	2370.2	-17.7 ± 0.4	-43.2 ± 0.4	25.5 ± 33
Apt09	-102.2	2134.4	2.8 ± 0.4	-54.2 ± 0.5	56.9 ± 17
Apt10	-104.8	2574.4	-44.3 ± 0.5	-96.1 ± 0.9	51.8 ± 14
Apt01rc	-94.7	1538.5	-32.1 ± 0.3	-43.5 ± 0.3	11.4 ± 36
Apt02rc	-103.5	1844.9	-60.0 ± 0.5	-61.1 ± 0.5	1.1 ± 18
Apt03rc	-104.3	2086.9	-38.5 ± 0.4	-68.7 ± 0.4	30.2 ± 20
Apt04rc	-145.6	3215.5	-51.5 ± 0.5	-160.8 ± 0.5	109.2 ± 16
Apt05rc	-102.2	1314.1	-21.2 ± 1.1	-34.6 ± 1.8	13.4 ± 29
Apt06rc	-92.5	1799.6	-78.0 ± 0.6	-101.5 ± 0.4	23.5 ± 23
Apt07rc	-85.5	1404.5	-34.8 ± 0.5	-57.4 ± 0.5	22.6 ± 21
Apt08rc	-110.2	1850.3	-72.2 ± 0.4	-112.4 ± 0.4	40.2 ± 17
Apt09rc	-113.4	1961.9	-42.0 ± 0.3	-62.1 ± 0.3	20.1 ± 16
Apt10rc	-100.5	2019.3	-2.2 ± 0.3	-28.5 ± 0.3	26.3 ± 13

4.1.3.4. Refinement of CatB-Aptamer complexes and binding free energy calculations

To improve the final ranking of the docked complexes, they were further refined through 100 ns long MD simulations and binding free energy calculations using the MM-GBSA method. Similar strategies have been successfully applied before to enhance the ranking reliability after molecular docking (190,461,463,465). The results regarding the MD simulations of the docked complexes are presented in **Supplementary Figures S5–S10**. The structure of CatB was very stable throughout all the simulations. In contrast, only the complexes with the aptamers Apt01, Apt04, Apt07, Apt06rc, and Apt08rc presented a relatively low (i.e., around 10 Å) and stable RMSD of all non-hydrogen atoms. This difference can be attributed to the inherently high flexibility of the ssDNA molecule compared with the globular CatB protein. When analyzing the representative structure of each complex, it is observed that the aptamers Apt01, Apt04, Apt07, Apt06rc, and Apt08rc

have a closer interaction with CatB, wrapping a larger region of the protein (**Supplementary Figures S9 and S10**), consistent with the high buried surface area observed during the last 50 ns of the MD simulations (**Table 8**, 6th column). Moreover, these aptamers also present a high number of established base pairs, corroborated by the high number of intramolecular Watson-Crick hydrogen bonds, particularly in the case of Apt04.

Table 8. Intra- and intermolecular interactions of each aptamer. The number of base pairs, Watson–Crick (W.C.) hydrogen bonds, intermolecular hydrogen bonds, intermolecular salt bridges, and buried surface area were calculated for the last 50 ns of the MD simulations of the CatB-Apt docked complexes. The base pairs, hydrogen bonds, and salt bridges were presented as mean $\pm$ SD. The buried surface area is presented in $\text{\AA}^2 \pm$ standard deviation (SD). The aptamer chosen for experimental validation is highlighted in bold.

Aptamer	Intramolecular (Apt-Apt)		Intermolecular (CatB-Apt)		
	# Base Pairs	# W.C. H-Bonds	# H-Bonds	# Salt Bridges	Buried surface area ($\text{\AA}^2$)
Apt01	16.7 $\pm$ 0.9	42 $\pm$ 2	6 $\pm$ 2	2 $\pm$ 1	1755 $\pm$ 204
Apt02	17.5 $\pm$ 0.6	42 $\pm$ 1	6 $\pm$ 2	3 $\pm$ 1	1060 $\pm$ 171
Apt03	15.2 $\pm$ 0.6	26.6 $\pm$ 0.6	8 $\pm$ 2	2 $\pm$ 1	1328 $\pm$ 207
Apt04	22.1 $\pm$ 0.8	49 $\pm$ 1	7 $\pm$ 2	3 $\pm$ 1	1422 $\pm$ 145
Apt05	18.0 $\pm$ 0.1	43.7 $\pm$ 0.6	4 $\pm$ 1	2 $\pm$ 1	824 $\pm$ 60
Apt06	17.8 $\pm$ 0.9	40 $\pm$ 2	5 $\pm$ 2	3 $\pm$ 1	1556 $\pm$ 271
Apt07	15.4 $\pm$ 0.6	32 $\pm$ 1	7 $\pm$ 2	2 $\pm$ 1	2094 $\pm$ 158
Apt08	16.7 $\pm$ 0.5	41 $\pm$ 1	3 $\pm$ 1	1 $\pm$ 1	664 $\pm$ 157
Apt09	17.4 $\pm$ 0.6	37 $\pm$ 2	5 $\pm$ 2	2 $\pm$ 1	850 $\pm$ 166
Apt10	17.7 $\pm$ 0.9	38 $\pm$ 1	5 $\pm$ 2	2 $\pm$ 1	1527 $\pm$ 371
Apt01rc	15.6 $\pm$ 0.5	38.8 $\pm$ 0.5	3 $\pm$ 2	2 $\pm$ 1	687 $\pm$ 76
Apt02rc	10.9 $\pm$ 0.6	27 $\pm$ 2	8 $\pm$ 2	3 $\pm$ 1	1082 $\pm$ 92
Apt03rc	10.2 $\pm$ 0.8	20 $\pm$ 1	6 $\pm$ 2	3 $\pm$ 1	1177 $\pm$ 121
Apt04rc	11 $\pm$ 1	26 $\pm$ 2	7 $\pm$ 2	1 $\pm$ 1	2184 $\pm$ 152
Apt05rc	17.0 $\pm$ 0.2	43.7 $\pm$ 0.6	3 $\pm$ 1	1 $\pm$ 1	1091 $\pm$ 63
Apt06rc	14.8 $\pm$ 0.7	38 $\pm$ 2	9 $\pm$ 3	1 $\pm$ 1	1589 $\pm$ 120
Apt07rc	10.9 $\pm$ 0.3	29 $\pm$ 1	4 $\pm$ 1	1 $\pm$ 1	881 $\pm$ 59
Apt08rc	11.2 $\pm$ 0.5	28.9 $\pm$ 0.7	10 $\pm$ 3	4 $\pm$ 2	1957 $\pm$ 126
Apt09rc	11.4 $\pm$ 0.6	27 $\pm$ 2	4 $\pm$ 1	2 $\pm$ 1	981 $\pm$ 81
Apt10rc	16.0 $\pm$ 0.2	42.6 $\pm$ 0.6	1 $\pm$ 1	0 $\pm$ 0	401 $\pm$ 135

An analysis of the hydrogen bonds established between CatB and the aptamers showed that several amino acids are involved in at least one hydrogen bond with an aptamer for

more than 50% of the simulation time (**Table S2**), with Thr₁₃₉ being the one that appears more in the interactions. Moreover, the amino acids involved in the top 5 longest hydrogen bonds established between CatB and the aptamers are presented in **Table S1**.

The binding free energy calculations predicted by the MM-GBSA method showed that all CatB-aptamer complexes herein generated have a favourable total ΔG_{bind} , except for the complex with Apt09. Nonetheless, considering the best total ΔG_{bind} values, three aptamers stand out: Apt04, Apt06rc, and Apt08rc, with -71.5 ± 0.5 , -78.0 ± 0.6 , and -72.2 ± 0.4 kcal.mol⁻¹, respectively (**Table 7**, 4th column). Interestingly, these complexes are among those displaying a higher number of hydrogen bonds and salt bridges between the protein and the aptamer (**Table 8**, 4th and 5th columns). Considering the different energetic terms, non-polar energies favour the binding over polar energies. In order to select an aptamer for experimental validation, both the Haddock score and the total binding free energy were considered (**Figure 53**). Since the total ΔG_{bind} is generally taken as a good indication of the binding affinity in a complex (460,461,466), it was first used to select the best aptamers (Apt04, Apt06rc, and Apt08rc). From those, we picked Apt08rc as it has a considerably lower Haddock score (-110.2 vs -90.3 for Apt04, and vs -92.5 , for Apt06rc).

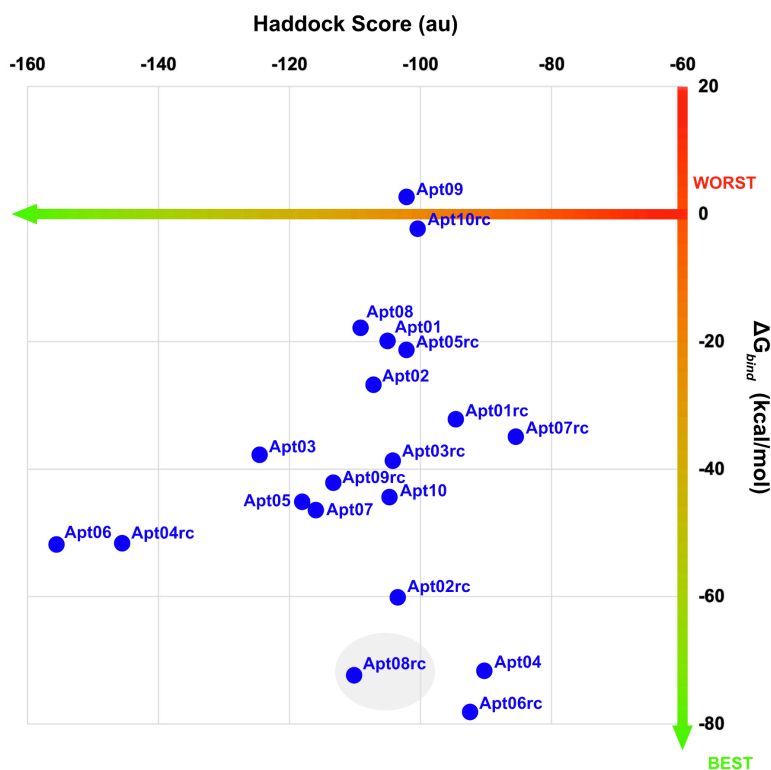


Figure 53. Comparison of the Haddock score and total binding free energy (in kcal.mol⁻¹) obtained using the MM-GBSA method for the 20 aptamers under study. The aptamer that was tested experimentally is highlighted with a grey circle (Apt08rc).

A detailed analysis of the Apt08rc and CatB complex revealed that the interaction between this aptamer and CatB occurs mainly through electrostatic interactions. Considering all the residues within 3.5 Å from the Apt08rc, it is clear that the interface between CatB and Apt08rc consists mainly of positively charged residues, with a few polar residues with uncharged side groups, one non-polar residue, and one negatively charged one, as seen on **Figure 54A** and **B**. Of note, the near absence of negatively charged residues is expected due to the polyanionic nature of DNA as a result of the negatively charged phosphate groups that are part of its backbone. The RMSD analysis of the MD simulation of the CatB-Apt08rc complex is depicted in **Figure 54C** and **D** and shows that the complex is relatively stable throughout the simulation, with no significant changes in conformation, with an RMSD no higher than 10 Å, contrasting with some of the other aptamers tested (**Supplementary S5–S8**). Moreover, as seen in **Figure 54D**, the aptamer seems to be the major contributor for a high RMSD value since it accounts for the majority of the RMSD value of the entire complex, while the CatB protein maintains its RMSD low throughout the simulation, meaning that its overall structure is well maintained.

The number of hydrogen bonds and salt bridges established between CatB and the Apt08rc aptamer is the highest among all the 20 aptamers (**Table 8**, 4th and 5th columns), slightly increasing during the MD simulation of the complex, as seen in **Figure 55**, reaching a mean of 10 hydrogen bonds and 4 salt bridges. The longest hydrogen bond established between CatB and all the aptamers occurs between the amino acid Ser₆₅ and the Apt08rc, with a maximum lifetime of 31.0 ns. This hydrogen bond is established for 64% of the simulation time (**Table S1**).

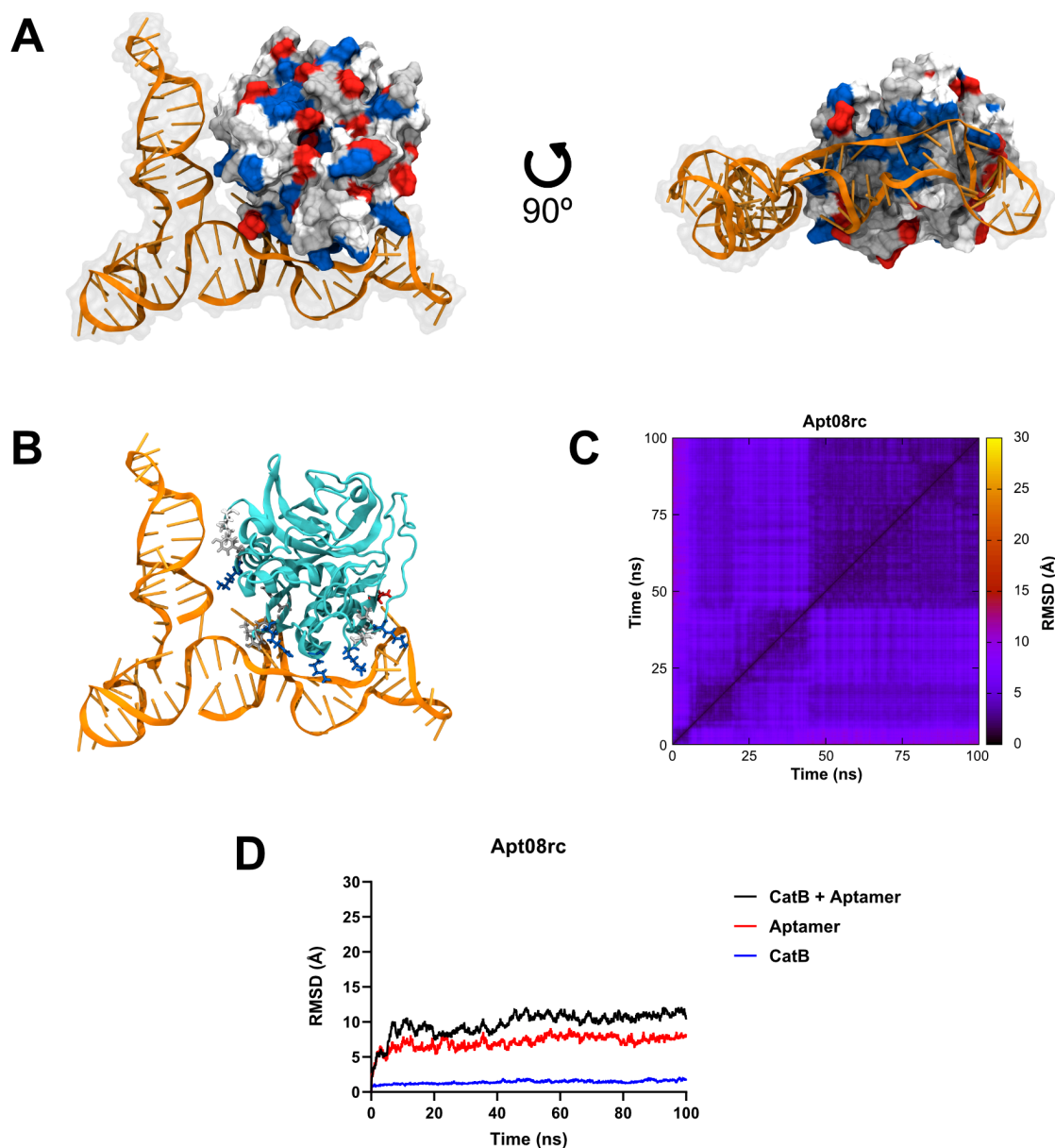


Figure 54. Analysis of the molecular dynamics simulation of the CatB-Apt08rc complex. (A) Representative structure of the CatB-Apt08rc complex obtained by K-means clustering of the molecular dynamics simulation observed from the side (left) and from below (right). CatB is represented in surface and coloured according to its residue type. Blue, positively charged side groups; red, negatively charged side groups; white, non-polar with aliphatic or aromatic side groups; Grey, polar with uncharged side groups. (B) Representative structure of the CatB-Apt08rc complex with CatB represented in NewCartoon, with all the residues within 3.5 Å from the Apt08rc represented in Licorice and coloured according to its residue type. (C) Pairwise RMSD of the trajectory of the CatB-Apt08rc complex. The colour scale is represented on the right going from a dark purple (0 Å) to yellow (30 Å). (D) RMSD representation of either the entire complex, Apt08rc, or CatB as a function of time for the 100 ns of the simulation.

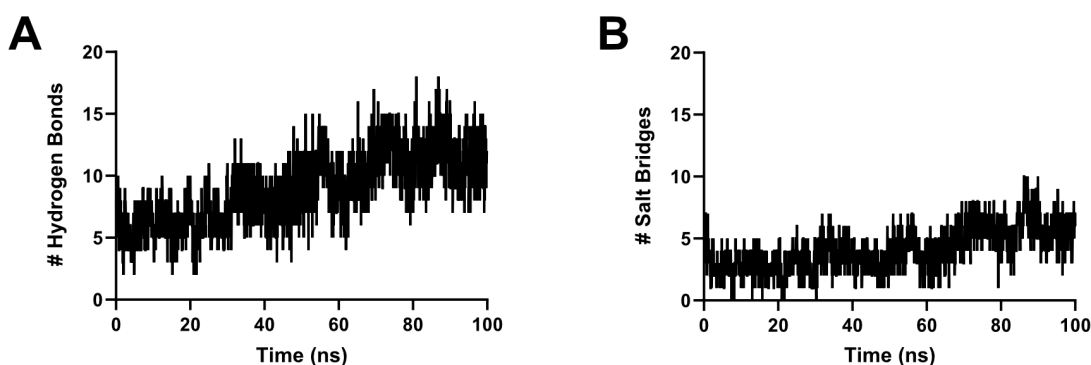


Figure 55. Hydrogen bonds and salt bridges established between CatB and Apt08rc. (A) Number of hydrogen bonds established between the CatB and Apt08rc, as a function of time for the 100 ns of the MD simulation. (B) Number of salt bridges established between CatB and Apt08rc, as a function of time for the 100 ns of MD simulation.

Taking these results into account, both Apt01 (aptamer with the highest number of copies after 8 rounds of selection) and Apt08rc (aptamer with the best performance in the computational study) were chosen for further experimental validation.

4.1.3.5. Calculation of the aptamer's affinity properties

First, calibration curves and kinetic assays were conducted for the two selected aptamers to evaluate the fluorescence signal change and to calculate the K_d values, respectively. Apt08rc displayed a linear fluorescent signal with the increase of aptamer concentration (**Figure 56A**). This result allowed to choose a suitable concentration for the kinetic assay, where Apt08rc was fixed at 50 nM, and CatB's concentration increased up to 100 nM (**Figure 56B**). The same assessment was performed for Apt01, and **Figure 56C** and **D** show the results for the calibration curve and kinetics assay, respectively, of this aptamer in solution.

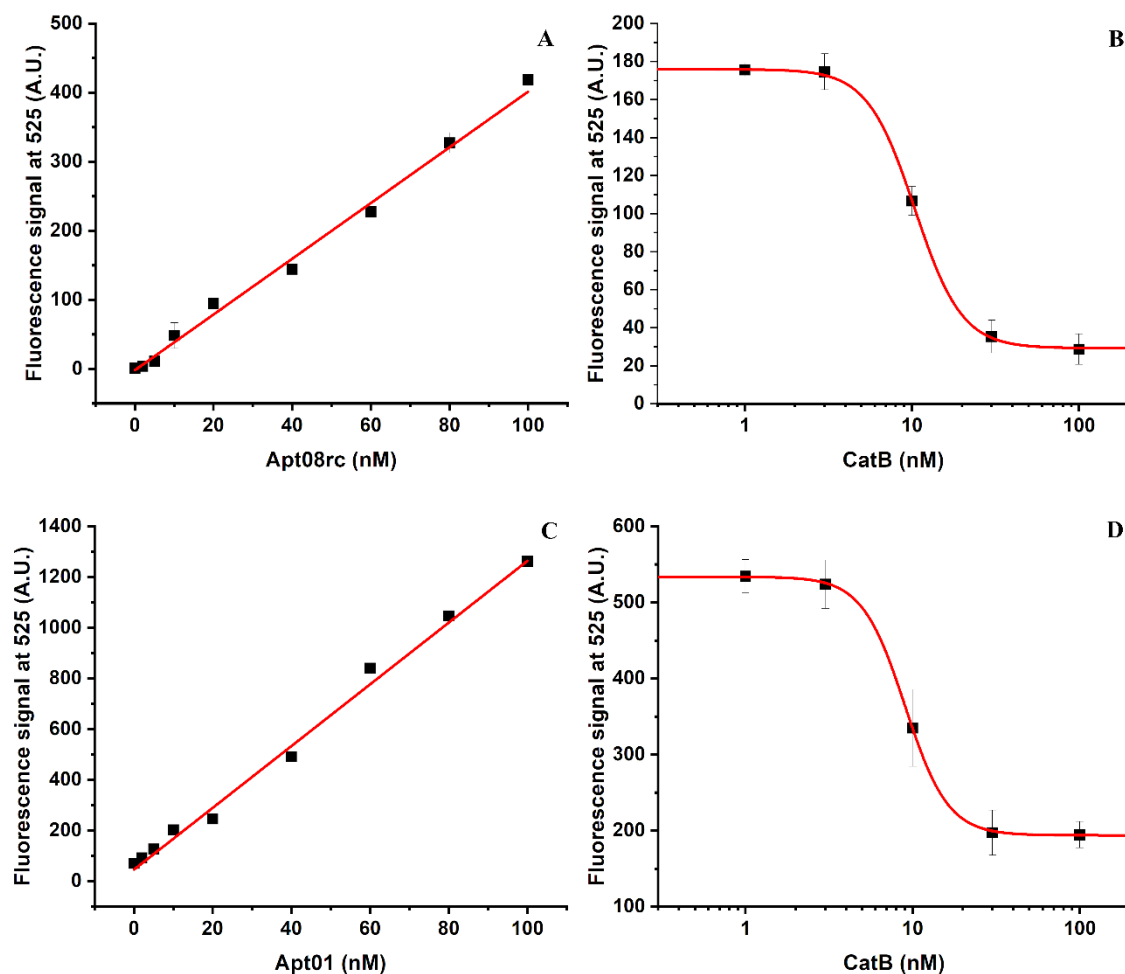


Figure 56. Calibration curve (A) and kinetic assay (B) for Apt08rc, and calibration curve (C) and kinetic assay (D) for Apt01. Experiments were conducted in duplicate, and results correspond to the mean $\pm$ standard deviation.

Regarding the calibration curves, for Apt08rc, the linear regression returned $y = (4.0 \pm 0.1)x - (1.8 \pm 1.8)$ with a correlation factor > 0.98 and for Apt01, the linear regression returned $y = (12.2 \pm 0.4)x - (46.3 \pm 17.5)$ with a correlation factor > 0.99 . Fitting to Langmuir isotherm, the calculations were determined using the model Hill 1 equation $y = \text{start} + \frac{(\text{end} - \text{start}) \times x^n}{k^n + x^n}$, with $\text{start} = 175.95 \pm 1.24$, $\text{end} = 29.23 \pm 2.67$ and $n = 3.07 \pm 0.57$ for Apt08rc and $\text{start} = 533.79 \pm 23.95$, $\text{end} = 193.86 \pm 18.03$ and $n = 3.38 \pm 0.75$ for Apt01. The calculated apparent K_d for Apt08rc was 10.4 ± 0.3 nM, and for Apt01 was 9.0 ± 2.1 nM, showing both high affinity towards the CatB.

4.1.3.6. Binding assays of the selected aptamers against CatB-overexpressing prostate cancer cells

SELEX is a successful technique to generate specific oligonucleotide sequences towards a purified protein. However, it is well known that the selected aptamers may not recognize the same protein in conditions different from those selected or in their native conformation in a cell (467). Therefore, we investigate whether Apt01 and Apt08rc possess the ability to recognize CatB using a prostate cell line (PC-3), which is described in the literature as a cell line with secretion of CatB at the cellular membrane (468–470). PC-3 cells were incubated with increasing concentrations of FAM-labelled Apt01 or Apt08rc for 1 h at room temperature. The initial library was used as a negative control.

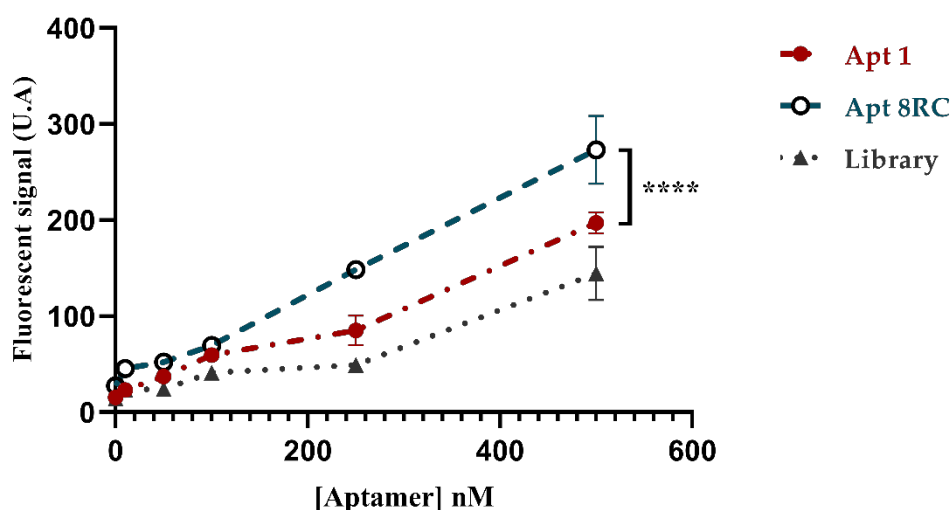


Figure 57. Binding assays with PC-3 cells (cellular density of 1×10^4 cells per well) and increased concentrations of library (grey), Apt01 (red), and Apt08rc (blue) at room temperature. **** $p < 0.0001$.

Figure 57 shows that, for higher concentrations of aptamer, 250 nM and 500 nM, Apt08rc displays a significantly higher affinity for PC-3 cells in comparison with Apt01 ($p < 0.0001$). Moreover, the fluorescent signal difference between the library and each aptamer also showed statistical significance, with a p -value < 0.005 for Apt01 and a p -value < 0.0001 for Apt08rc, confirming the enrichment of the selected aptamers to the target CatB. These results corroborate the bioinformatics data, where Apt08rc showed a much more favourable ΔG_{bind} than Apt01 (-72.2 ± 0.4 vs. -19.8 ± 0.6).

As previously mentioned in the Introduction section, aptamers represent a valuable tool with high potential in cancer therapy. The use of SELEX, combined with the prediction

competence of the *in silico* tools, provides a powerful approach for the selection of aptamers with high affinity towards a target of interest. The Apt08rc, as the preliminary binding assays suggest, shows increased affinity when compared to Apt01 or the library, recognizing CatB-overexpressing PC-3 cells.

4.1.4. Conclusion

This work identifies for the first time an aptamer that can efficiently bind to CatB-overexpressing cancer cells. To our knowledge, this is the first aptamer identified for this protein. SELEX, combined with NGS, allowed the identification of 10 aptamer candidates that, along with their respective reverse complementary sequences, were further studied using bioinformatics tools. *In silico* predictions pointed to Apt08rc as the best aptamer candidate, then chosen for further characterization, along with aptamer Apt01. Kinetic studies provided similar results for both aptamers with free purified CatB, with apparent K_d values of 10.4 ± 0.3 nM and 9.0 ± 2.1 nM for Apt08rc and Apt01, respectively. Binding studies with the PC-3 cells, known to produce high levels of CatB at the surface, confirmed the bioinformatics predictions as Apt08rc was more efficient in binding PC-3 cells than Apt01. Future studies should aim to fully characterize the selected aptamer and validate it with other cell lines. Ultimately, this novel ligand could be a helpful tool for CatB monitoring in cancer diagnosis, as growing evidence supports this protein as a predictive biomarker in cancer progression or for targeting nanoformulations for CatB-overexpressing cancer cells.

D. FINAL REMARKS

This page was intentionally left blank.

This PhD thesis is entitled “Innovative approach to fight tuberculosis and malaria targeting the extraordinary PLP synthase macromolecular complex.” However, a PhD is not a sprint; it is more akin to hiking, where there are ups and downs, and sometimes detours must be taken to reach the final destination. That is how I felt regarding this PhD project. Even though it was not possible to complete all the defined milestones (i.e., fully characterize the catalytic mechanism of PLP synthase and its interactions), other objectives were achieved. As in hiking, it is not the destination that matters, but the journey along the way. As such, this PhD project was grounded in two main pillars: proteins and DNA. Moreover, as this work is based on computational methodologies, a brief discussion about the advantages and limitations of the methods used is also presented at the end of this chapter.

1. Proteins

Regarding the study of proteins, particularly enzymes, the focus was **pyridoxal 5'-phosphate** (PLP). Here the focus resided on two aspects of PLP, its **synthesis** and its **role as a cofactor**.

1.1. Pdx2 – a key enzyme in PLP biosynthesis

Starting with the synthesis of PLP, the focus was PLP synthase, the sole enzyme required for the *de novo* synthesis of PLP, the ubiquitous pathway present in most life forms, especially pathogens like *Plasmodium falciparum* and *Mycobacterium tuberculosis*. However, there is a notable exception. Most mammals, in particular humans, cannot synthesize PLP, resorting to acquiring it from their diets in the form of other B₆ vitamers. This makes PLP synthase a promising drug target against diseases like Malaria and Tuberculosis caused by the pathogens mentioned above. In short, PLP synthase is a 24-meric complex composed of 12 synthase subunits, Pdx1, and 12 glutaminase subunits, Pdx2. Only when fully assembled is this complex capable of synthesizing PLP *in vivo*. Therefore, the two key aspects regarded as having the best inhibitory potential are interfering with the catalytic mechanism of Pdx1, specific to this enzyme, or hindering complex formation. Unfortunately, neither Pdx1's catalytic mechanism nor the residues contributing the most to Pdx1:Pdx1 and Pdx1:Pdx2 interfaces could be characterized during this PhD thesis. Nevertheless, they are two relevant topics that should be addressed in the future.

However, the same is not true regarding Pdx2, whose catalytic mechanism was thoroughly studied. Pdx2 is the sole provider of ammonia to the Pdx1 *in vivo*, and without it, the catalytic mechanism cannot proceed, so its importance must not be neglected.

Furthermore, the fact that it employs a non-canonical Cys-His-Glu catalytic triad for its catalysis highlights the importance of its study even more, as contrary to Ser-His-Asp/Glu catalytic triads, whose catalytic mechanism is well established, there is less information about it. Therefore, due to this work, the following general conclusions could be taken about the Cys-His-Glu catalytic triad, whose mechanism is very similar but not equal to other catalytic triads such as Ser-His-Asp/Glu or Ser-(cis)Ser-Lys, mostly because it employs a different nucleophile:

- The catalytic process involves two main stages:
 - i) Formation of an acyl-enzyme intermediate (i.e., adduct) in which the enzyme becomes covalently bonded to the enzyme.
 - ii) Hydrolysis of the adduct resulting in the reaction product's formation and enzymatic turnover.
- In the process of forming the acyl-enzyme intermediate, it is typically observed the formation of an anionic tetrahedral reaction intermediate upon the nucleophilic attack of the Cys/Ser. However, this is only possible due to the existence of an oxyanion hole that stabilizes the negative charge that forms in the oxygen of the carbonyl group. The oxyanion hole is typically formed by the backbone amino group of some residues in the enzyme's active site.
- The catalytic Glu does not have a direct key role in the mechanism (i.e., it does not receive a proton from the catalytic His) as observed in other Ser-His-Asp/Glu enzymes. Nonetheless, Glu is required to maintain the correct alignment and orientation of His imidazole ring in relation to Cys and the substrate.

However, due to the nature of the cysteine residue, whose side chain has a pK_a of 8-9, there are also some key differences:

- The first step of the catalytic process of an enzyme employing a Cys-His-Glu catalytic triad is the activation of the Cys residue. This involves the abstraction of its proton by the catalytic His, forming a $\text{Cys}^-:\text{His}^+$ ionic pair. This is required since only the thiolate form of Cys (R-S^-) is a good enough nucleophile. Contrary, in other catalytic triads, such as the ones that employ Ser as the nucleophile, the proton abstraction is concomitant with the nucleophilic attack (50).

- No tetrahedral reaction intermediary was observed during the hydrolytic stage. Contrary to what is suggested for Ser- catalytic triads, the nucleophilic attack of a water molecule to the acyl intermediate resulted in the immediate collapse of the structure, with the reformation of the Cys⁻:His⁺ ionic pair.
- Lastly, the rate-limiting step of this catalytic process was the hydrolysis (which requires 18.2 kcal.mol⁻¹ and agrees very well with the experimental k_{cat} of ~19 kcal.mol⁻¹), contrarily to other enzymes employing the catalytic triads Ser-His-Asp/Glu (e.g., Subtilisin (471)) or Ser-(cis)Ser-Lys (e.g., Malonamidase E2 (307)) whose rate-limiting step is the first one (i.e., formation of the tetrahedral reaction intermediate).

All the collected information can be used to develop new drugs (i.e., inhibitors) targeting Pdx2. Due to the nature of the catalytic process, the inhibitor could either be reversible or irreversible. In the case of an irreversible inhibitor, it would have to react more favourably with the enzyme than its natural substrate, glutamine, i.e., it should have a lower Gibbs free energy of activation and be very exergonic. This would form a reaction intermediate that would block the catalytic cycle by making it virtually impossible for the reaction to proceed or for the reverse reaction to occur. On the other hand, a competitive, reversible inhibitor would only have to compete with the natural substrate for the enzyme's active site, preferably having a lower K_M . Since Pdx2 is not present in humans, this could seem a good strategy; however, its similarities with other glutaminases could raise problems by off-target toxicity.

1.2. XometC – role of PLP as a cofactor

Moreover, apart from Pdx2, the catalytic mechanism of a PLP-dependent enzyme was also studied, highlighting the importance and versatility of PLP as a cofactor. The enzyme in question is XometC, present in the pathogenic bacteria *Xanthomonas oryzae pv. oryzae*, commonly abbreviated as Xoo, which is the causing agent of a destructive disease that affects rice crops worldwide, having a substantial environmental and financial impact. As such, XometC has often been regarded as a potential drug target in developing new bactericides against this disease. XometC is a promiscuous enzyme with two catalytic activities, cystathionine γ -lyase and L-serine ammonia-lyase. Our studies focused their attention on the latter, a β -elimination reaction, where the hydroxyl group of the substrate L-serine is eliminated in the form of water, with the formation of 2-aminoacrylate.

This study entailed the complete catalytic mechanism, from the formation of the external aldimine with the substrate until the enzymatic turnover, with the product's release and re-formation of the internal aldimine. From this, some key points were observed regarding XometC's L-serine ammonia-lyase activity:

- The entire catalytic process requires 14 sequential steps since the attack of the substrate to the internal aldimine, until the formation of the product of the reaction and enzymatic turnover.
- The rate-limiting step of this catalytic process is the result of the proton abstraction by the catalytic Lys immediately followed by the β -elimination of the hydroxyl group of L-serine. This has a total activation free energy barrier of $19.1 \text{ kcal.mol}^{-1}$, which closely agrees with the experimental kinetic data available, whose activation barrier is $18.9 \text{ kcal.mol}^{-1}$ (72).
- The conversion of 2-aminoacrylate into the final product, pyruvate, is impossible to occur inside the enzyme's active site. Nevertheless, we have shown that it can occur in the cytosol with a rate-limiting step whose activation barrier ($22.6 \text{ kcal.mol}^{-1}$) closely agrees with the experimental data described in the literature (21 kcal.mol^{-1}) (342).
- Lastly, we observed that the dissociated product, 2-aminoacrylate, can work as a proton shuttle between the phosphate group and the catalytic Tyr, reestablishing enzymatic turnover.

In a more generalized manner, this study also allows us to understand better some key roles played by the PLP cofactor in most PLP-dependent enzymes. It was found that each structural feature of the PLP cofactor plays a key and different role in different stages of the catalytic process. Some of the most important are described below:

- The phosphate group of PLP plays a key role in anchoring the cofactor to the enzyme's active site. Nevertheless, its role goes beyond that, also working as a base, removing a proton from catalytic Lys, activating it to form the second gem-diamine, ultimately leading to product release and enzymatic turnover. As such, PLP's phosphate group also has a role in proton storage during the catalytic process.
- Oxygen O3 of the PLP's pyridine ring works as a proton shuttle between the catalytic Lys and the amino group of the substrate/product when in the gem-diamine form.

- A positively charged residue, Arg₃₇₄ in XometC, guarantees the correct orientation and alignment of the substrate in the active site in relation to the PLP cofactor. Furthermore, by interacting with the carboxylate group of the substrate, Arg₃₇₄ constrains the bond that will be cleaved perpendicularly to the plane of the PLP cofactor, which is fundamental for catalysis to occur, as proposed by the Dunathan hypothesis.
- This is only possible due to the role played by the pyridinium ring of PLP in delocalizing the excess negative charge formed during bond cleavage across the cofactor. This delocalization is favoured by the presence of a negatively charged residue below the pyridinium ring of PLP, Asp₁₈₅ in XometC.
- The free energy profile reveals that the external aldimine formation is very favourable, suggesting that there may be a shift towards this structure under physiological conditions.

Furthermore, since XometC is important for Xoo virulence, all the gathered knowledge about its catalytic process can be used to develop novel bactericides targeting it (308). As previously mentioned, these inhibitors can be either reversible or irreversible. In the case of an irreversible inhibitor, it must be designed to become covalently bonded to the PLP cofactor, efficiently inactivating the enzyme. The best inhibitor would result in an extra stabilization of the quinonoid reaction intermediate, impeding the reaction from proceeding either backwards, towards the external aldimine, or forward, resulting in the reaction's product.

This page was intentionally left blank.

2. DNA

Regarding DNA, two collaborations provided the opportunity to study its role besides storing genetic information. DNA is a highly flexible molecule capable of acquiring different conformations (i.e., non-B DNA conformations) and selectively interacting with other proteins (i.e., DNA aptamers), which makes them promising candidates as new therapeutic agents against several diseases, with cancer as the prominent example. This also constituted an opportunity to acquire proficiency in the application of molecular dynamics simulations, particularly in the study of nucleic acids in a highly emerging area of development.

When we hear the word DNA, we are often led to think of the double-helix B-DNA structure as initially described by Watson and Crick, based on the work of Rosalind Franklin and Raymond Gosling, among others. However, that is not the only conformation that DNA can acquire. We now know that although B-DNA is the most common form under physiological conditions, DNA can also acquire other conformations, such as hairpins and G-quadruplexes, collectively known as non-B DNA conformations. Although their roles are still not completely understood, there is evidence that non-B DNA conformations may have implications in cellular functions, such as transcription regulation. However, these conformations have also been linked with diseases due to increased replication errors, repair errors, and genomic instability in regions of the DNA where these structures are more prone to form. A bioinformatic analysis of the mtDNA found a 93-nt stretch of DNA more prone to suffer alterations, which they named Structure A. This work was a proof of concept, aiming to study the stability of Structure A compared to the conformation acquired by other 93-nt mtDNA stretches. The rationale was that if Structure A non-B DNA conformation was more stable than the rest, then its 3D structure must be related to the mutagenic instability that it causes. With this work, we developed a pipeline that allows us to consistently predict the 3D conformation of any stretch of ssDNA/RNA, simulate it under physiological conditions, and analyze those simulations. Using this workflow, our results showed that the conformation acquired by Structure A is highly stable and presents low flexibility, contrary to the rest of the structures studied.

However, DNA is not only a naturally occurring molecule, be it B- or non-B DNA. DNA can be synthesized with the length and nucleotide composition of our liking. That has been the idea behind synthesizing DNA aptamers, DNA molecules that are selected for their high affinity towards a specific target, therefore promising in the development of target therapies against several diseases, especially tumour cells. However, synthesizing and selecting those aptamers, a process known as SELEX, can be very time-consuming and

expensive, especially as one selects between better and better aptamers. For that reason, *in silico* methods have been regarded as a useful tool helping to identify the best ones. In collaboration with a research group from the University of Minho, Braga, Portugal, we aimed to select the aptamer with the highest affinity towards Cathepsin B, a protein overexpressed in several cancers, especially prostate cancer. Our part in the collaboration was to, using computational methods, identify the best from 20 aptamers previously obtained with SELEX. Therefore, by making some modifications to the workflow used in the non-B DNA conformations prediction, this was achieved with great success, with the *in silico* prediction corroborated experimentally. Moreover, this workflow cannot only predict the best aptamers but also rationalize their affinity by identifying where they bind to the target and the type of interactions they make. These data can then be used to improve the aptamers even further.

3. Methodology

3.1. Model preparation

Computational methods are invaluable for describing chemical reactions with atomistic detail since, even nowadays, it is still almost impossible to accomplish this task using experimental methodologies. In any case, we must have a 3D structure of our system (e.g., protein, DNA, or ligand) to which those methods can be applied. Unfortunately, until a few years ago, we were entirely dependent upon the availability of X-ray crystallography, NMR, or cryo-EM structures. When those were unavailable, we had to resort to homology modelling, with all the problems it may entail. However, with the rise of artificial intelligence-based methods, modelling enzyme structures have become increasingly reliable due to tools like AlphaFold.

Moreover, in the case of nucleic acids, there are not many 3D structures experimentally determined, especially of single-stranded DNA (ssDNA) or RNA, so we must resort to their prediction. However, while tools like UNAFold can predict the secondary structure of DNA and RNA, the same does not apply to 3D structures, which currently only have tools available for the prediction of RNA. Therefore, our only option is to generate a 3D RNA structure and then manually convert it to an analogous 3D ssDNA structure, with all the oversimplifications and assumptions it may entail. Nonetheless, in the case of proteins, most 3D structures still result from X-ray crystallography, as has been done for decades. The only problem with X-ray structures is that the conditions needed for crystal formation are often entirely different from physiological conditions, which may lead to a 3D structure that does not resemble the one found in natural conditions. Moreover, special attention must be given to structures with poor resolution, especially regarding ligands, as those may be incorrectly modelled into the density map, leading to incorrect structures and, ultimately, incorrect conclusions (9).

3.2. Molecular mechanics

Molecular mechanics methods like molecular dynamics (MD), molecular docking, and binding free energy calculations (using the MM-GBSA approach) were employed throughout this PhD project, mainly in research involving DNA molecules. Most force fields commonly used in biomolecular simulations, like the AMBER ff14SB for proteins, or the bsc1 for nucleic acids, are additive, as we have seen in **Chapter B**. Although they have been successfully employed in many works throughout the years, they present some

known deficiencies, primarily regarding the description of electrostatic interactions, an aspect of particular importance in the case of a highly polyanionic molecule like DNA. To solve this problem, polarizable force fields specially designed for DNA have been developed (472). Nevertheless, they still present two main problems: i) they are much more computationally demanding than additive force fields, and ii) related to the first point, there are not enough studies employing these force fields which would truly solidify their claims. Do not forget, “No matter how effective or powerful a new software solution is, it is only as good as the number of people using it.” — Philip Kushmaro, CIO

Moreover, the exploration of the molecular conformational space of highly flexible molecules like ssDNA structures is very limited. In an attempt to fully explore every possible conformation, we would have to extend our simulations closer to the second scale, which is simply unfeasible. Other methods, such as Monte Carlo simulations or Metadynamics simulations, could help, but even those do not allow the full exploration of the conformational space. As such, the methodology used in the works presented in this thesis seems the most sensible.

Regarding the protein-DNA docking studies, some questions can be raised. For many years the focus of molecular docking was centred on protein-small ligands. Then, more and more software started to surge for protein-protein docking. What happened was that most of that software was then adapted also to accept nucleic acids, as was the case of HADDOCK (172). By being an afterthought, one might argue that they were not prepared for such a use case. Nevertheless, several benchmarks and improvements have been made over the years, demonstrating that HADDOCK can deal with this type of system accurately (173,177,178,473).

The last point regarding molecular mechanics is the calculation of binding free energy calculations using the MM-GBSA approach. Although performing these calculations is already a huge step up in comparison with estimating the best complexes solely based on the docking score, there are still some approximations that should be considered, as the absence of the entropic contribution (not considered in our calculations since the approximation used for its determination is not very accurate), or the low accuracy of the SASA model used for the calculation of the nonpolar contribution of the solvation energy (458). Nevertheless, in comparison with more rigorous methods, MM-GBSA is still the one with the best compromise between accuracy and computational cost.

3.3. Quantum mechanics

Despite all the computational power we have at hand, the study of an entire biomolecular system using *ab initio* methods is still not feasible. Nevertheless, some approaches allow us to balance speed and accuracy well, as is the case of the hybrid methodology, QM/MM, which led to the attribution of the Nobel Prize in Chemistry in 2013. Using QM/MM, one can separate a molecular system into the QM and MM regions. The chemical species most relevant for a particular reaction will be included in the QM region, while the rest is part of the MM region. However, one cannot be too greedy. For example, if we do not consider species relevant for the catalysis, even if they are not directly involved, our results may not be correct. On the other hand, if our QM region is too big, our calculations will not finish in a reasonable time frame. Therefore, our goal is to achieve a perfect balance between the size of the QM region and the computational cost it will require, which is a form of art on its own. Nonetheless, applying QM/MM methods makes the computational study of biomolecular systems relevant to health and the environment possible.

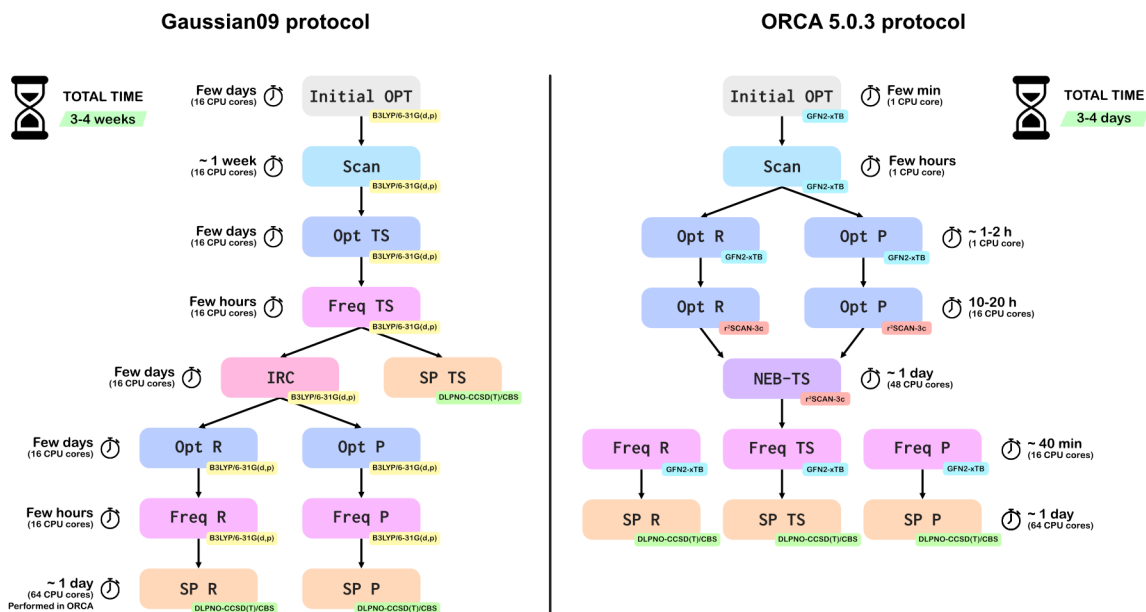
Moreover, the experimental data about a catalytic mechanism or the enzyme that performs it is often limited or nonexistent. In those situations, ensuring an adequate exploration of the reactional space is even more critical. This means that, in practice, several mechanistic hypotheses must be tested until we find the one most likely to occur *in vivo*. We know from experience that a wrong mechanistic proposal results in energy barriers incompatible with life as it is inside a cell. Therefore, we always aim to find the most kinetically and thermodynamically favourable proposal. Another alternative would be using QM/MM MD methods, which are often used to explore the conformational space of proteins and ligands more efficiently, however, at the expense of much higher computational costs (474). Therefore, people tend to use less accurate semi-empirical methods or lower-rung DFT functionals to describe the QM region to balance this cost out. This matter is still a question of debate among the scientific community, but for now, with the computational power we have available, we consider that using static QM/MM methods with a higher theoretical level leads to more accurate results.

This is especially true when we are fortunate to have experimental kinetic data at our disposal, which serves as an essential checkpoint to validate the energy values we computationally determined. However, we can only do so if high theoretical levels are used to calculate the final energies, as those values must be accurate. Consider the following example. The Gibbs free energy of activation determined for the reaction we consider to be the rate-limiting step is 18 kcal.mol⁻¹, which according to the Eyring

equation, at 37 °C, corresponds to a turnover number of 1.3 s⁻¹. Now imagine that the real Gibbs free energy of activation is not 18 but 17 kcal.mol⁻¹, only 1 kcal.mol⁻¹ lower. That means that the real turnover number of this reaction is 6.8 s⁻¹, which is a considerable difference. This is why the final energy values determined for every reaction step were obtained using the DLPNO-CCSD(T) approach, as mentioned in the methodological chapter. Using this methodology, we can reduce the energy error to less than 1 kcal.mol⁻¹ while maintaining an affordable computational cost.

Lastly, I would like to point out the fact that throughout this PhD project, novel methodologies have been developed by the scientific community. At first, as seen in the work regarding the catalytic mechanism of Pdx2, we used Gaussian 09 software, employing a subtractive QM/MM scheme (i.e., ONIOM), using the hybrid functional B3LYP together with the 6-31G(d,p) basis set to describe the QM region. However, in version 4.2, the free ORCA software implemented several desirable features, like a native QM/MM module and the quick and relatively accurate semi-empirical GFN2-xTB method. This new method allowed for quick geometry optimizations, ideal for an initial optimization or exploring the reactional space by performing several relaxed scans.

Then, with version 5.0, ORCA implemented r²SCAN-3c, a composite electronic-structure method tailor-made for structure optimization. Optimizations with r²SCAN-3c are not only faster than with B3LYP/6-31G(d,p) but also more accurate. This paradigm shift revolutionized the way we worked. Therefore, I developed a new protocol, used for the first time in the study of XometC, based on the advantages of GFN2-xTB and r²SCAN-3c (described in more detail in **Chapter B**, and summarized in **Figure 58**), which allowed the characterization of a reactional step in a fraction of the time of the original protocol. From personal experience, while before, it took weeks to characterize a single reactional step, now, using this new protocol, I can achieve it in under three days. With better hardware coming out every year and new, more efficient methods being developed at this rate, who knows what could be possible in the next four years?



This page was intentionally left blank.

E. REFERENCES

This page was intentionally left blank.

1. Crick, F. H. On protein synthesis. *Symp. Soc. Exp. Biol.* **12**, 138–163 (1958).
2. Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).
3. Bley, C. J. *et al.* Architecture of the cytoplasmic face of the nuclear pore. *Science* **376**, eabm9129 (2022).
4. Petrovic, S. *et al.* Architecture of the linker-scaffold in the nuclear pore. *Science* **376**, eabm9798 (2022).
5. Radzicka, A. & Wolfenden, R. A Proficient Enzyme. *Science* **267**, 90–93 (1995).
6. Vitamins: MedlinePlus Medical Encyclopedia. *MedlinePlus* at <<https://medlineplus.gov/ency/article/002399.htm>>
7. Mukherjee, T., Hanes, J., Tews, I., Ealick, S. E. & Begley, T. P. Pyridoxal phosphate: Biosynthesis and catabolism. *Biochim. Biophys. Acta - Proteins Proteomics* **1814**, 1585–1596 (2011).
8. di Salvo, M. L., Contestabile, R. & Safo, M. K. Vitamin B6 salvage enzymes: Mechanism, structure and regulation. *Biochim. Biophys. Acta - Proteins Proteomics* **1814**, 1597–1608 (2011).
9. Fernandes, H. S., Ramos, M. J. & Cerqueira, N. M. F. S. A. The Catalytic Mechanism of the Pyridoxal-5'-phosphate-Dependent Enzyme, Histidine Decarboxylase: A Computational Study. *Chem. – Eur. J.* **23**, 9162–9173 (2017).
10. Dick, T., Manjunatha, U., Kappes, B. & Gengenbacher, M. Vitamin B6 biosynthesis is essential for survival and virulence of Mycobacterium tuberculosis. *Mol. Microbiol.* **78**, 980–988 (2010).
11. Kappes, B., Tews, I., Binter, A. & Macheroux, P. PLP-dependent enzymes as potential drug targets for protozoan diseases. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **1814**, 1567–1576 (2011).
12. Kronenberger, T., Schetttert, I. & Wrenger, C. Targeting the vitamin biosynthesis pathways for the treatment of malaria. *Future Med. Chem.* **5**, 769–779 (2013).
13. Müller, I. B., Hyde, J. E. & Wrenger, C. Vitamin B metabolism in Plasmodium falciparum as a source of drug targets. *Trends Parasitol.* **26**, 35–43 (2010).
14. Reeksting, S. B. *et al.* Exploring inhibition of Pdx1, a component of the PLP synthase complex of the human malaria parasite Plasmodium falciparum. *Biochem. J.* **449**, 175–187 (2013).
15. Wrenger, C. & Müller, I. B. in *Apicomplexan Parasites Mol. Approaches Target. Drug Dev.* (ed. Becker, K.) 235–249 (Wiley, 2011). doi:10.1002/9783527633883.ch12
16. Kronenberger, T. *et al.* Vitamin B6-dependent enzymes in the human malaria parasite Plasmodium falciparum: a druggable target? *BioMed Res. Int.* **2014**, 108516 (2014).
17. Fitzpatrick, T. B. *et al.* Two independent routes of de novo vitamin B6 biosynthesis: not that different after all. *Biochem. J.* **407**, 1–13 (2007).
18. Mittenhuber, G. Phylogenetic analyses and comparative genomics of vitamin B6 (pyridoxine) and pyridoxal phosphate biosynthesis pathways. *J. Mol. Microbiol. Biotechnol.* **3** 1, 1–20 (2001).

19. Lam, H. M. & Winkler, M. E. Metabolic relationships between pyridoxine (vitamin B6) and serine biosynthesis in *Escherichia coli* K-12. *J. Bacteriol.* **172**, 6518–6528 (1990).
20. Yang, Y., Zhao, G., Man, T.-K. & Winkler, M. E. Involvement of the gapA- and epd(gapB)-Encoded Dehydrogenases in Pyridoxal 5'-Phosphate Coenzyme Biosynthesis in *Escherichia coli* K-12. *J. Bacteriol.* **180**, 4294–4299 (1998).
21. Banks, J. & Cane, D. E. Biosynthesis of vitamin B6: direct identification of the product of the PdxA-catalyzed oxidation of 4-hydroxy-l-threonine-4-phosphate using electrospray ionization mass spectrometry. *Bioorg. Med. Chem. Lett.* **14**, 1633–1636 (2004).
22. Cane, D. E., Hsiung, Y., Cornish, J. A., Robinson, J. K. & Spenser, I. D. Biosynthesis of Vitamin B6: The Oxidation of 4-(Phosphohydroxy)-l-threonine by PdxA. *J. Am. Chem. Soc.* **120**, 1936–1937 (1998).
23. Sprenger, G. A. *et al.* Identification of a thiamin-dependent synthase in *Escherichia coli* required for the formation of the 1-deoxy-d-xylulose 5-phosphate precursor to isoprenoids, thiamin, and pyridoxol. *Proc. Natl. Acad. Sci.* **94**, 12857–12862 (1997).
24. Laber, B., Maurer, W., Scharf, S., Stepusin, K. & Schmidt, F. S. Vitamin B6 biosynthesis: formation of pyridoxine 5'-phosphate from 4-(phosphohydroxy)-l-threonine and 1-deoxy-d-xylulose-5-phosphate by PdxA and PdxJ protein. *FEBS Lett.* **449**, 45–48 (1999).
25. Cane, D. E., Du, S., Robinson, J. K., Hsiung, Y. & Spenser, I. D. Biosynthesis of Vitamin B6: Enzymatic Conversion of 1-Deoxy-d-xylulose-5-phosphate to Pyridoxol Phosphate. *J. Am. Chem. Soc.* **121**, 7722–7723 (1999).
26. di Salvo, M. L., Yang, E., Zhao, G., Winkler, M. E. & Schirch, V. Expression, purification, and characterization of recombinant *Escherichia coli* pyridoxine 5'-phosphate oxidase. *Protein Expr. Purif.* **13**, 349–356 (1998).
27. Ehrenshaft, M. & Daub, M. E. Isolation of PDX2, a second novel gene in the pyridoxine biosynthesis pathway of eukaryotes, archaeobacteria, and a subset of eubacteria. *J. Bacteriol.* **183**, 3383–90 (2001).
28. Guédez, G. *et al.* Assembly of the Eukaryotic PLP-Synthase Complex from Plasmodium and Activation of the Pdx1 Enzyme. *Structure* **20**, 172–184 (2012).
29. Fitzpatrick, T. B., Moccand, C. & Roux, C. Vitamin B6 Biosynthesis: Charting the Mechanistic Landscape. *ChemBioChem* **11**, 1185–1193 (2010).
30. Burns, K. E., Xiang, Y., Kinsland, C. L., McLafferty, F. W. & Begley, T. P. Reconstitution and Biochemical Characterization of a New Pyridoxal-5'-Phosphate Biosynthetic Pathway. *J. Am. Chem. Soc.* **127**, 3682–3683 (2005).
31. Zhang, X. *et al.* Structural insights into the catalytic mechanism of the yeast pyridoxal 5-phosphate synthase Snz1. *Biochem. J.* **432**, 445–454 (2010).
32. Ullah, N. *et al.* Solution Structures and Dynamic Assembly of the 24-Meric Plasmodial Pdx1-Pdx2 Complex. *Int. J. Mol. Sci.* **21**, (2020).
33. Strohmeier, M. *et al.* Structure of a bacterial pyridoxal 5'-phosphate synthase complex. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 19284–19289 (2006).

34. Bauer, J. A., Bennett, E. M., Begley, T. P. & Ealick, S. E. Three-dimensional Structure of YaaE from *Bacillus subtilis*, a Glutaminase Implicated in Pyridoxal-5'-phosphate Biosynthesis. *J. Biol. Chem.* **279**, 2704–2711 (2004).
35. Belitsky, B. R. Physical and enzymological interaction of *Bacillus subtilis* proteins required for de novo pyridoxal 5'-phosphate biosynthesis. *J. Bacteriol.* **186**, 1191–6 (2004).
36. Dong, Y. X., Sueda, S., Nikawa, J. & Kondo, H. Characterization of the products of the genes SNO1 and SNZ1 involved in pyridoxine synthesis in *Saccharomyces cerevisiae*. *Eur. J. Biochem.* **271**, 745–52 (2004).
37. Zhu, J., Burgner, J. W., Harms, E., Belitsky, B. R. & Smith, J. L. A New Arrangement of (β/α)₈ Barrels in the Synthase Subunit of PLP Synthase. *J. Biol. Chem.* **280**, 27914–27923 (2005).
38. Zein, F. *et al.* Structural Insights into the Mechanism of the PLP Synthase Holoenzyme from *Thermotoga maritima*. *Biochemistry* **45**, 14609–14620 (2006).
39. Flicker, K. *et al.* Structural and Thermodynamic Insights into the Assembly of the Heteromeric Pyridoxal Phosphate Synthase from *Plasmodium falciparum*. *J. Mol. Biol.* **374**, 732–748 (2007).
40. Gengenbacher, M. *et al.* Vitamin B6 biosynthesis by the malaria parasite *Plasmodium falciparum*: Biochemical and structural insights. *J. Biol. Chem.* **281**, 3633–3641 (2006).
41. Semmelmann, F., Hupfeld, E., Heizinger, L., Merkl, R. & Sterner, R. A Fold-Independent Interface Residue Is Crucial for Complex Formation and Allosteric Signaling in Class I Glutamine Amidotransferases. *Biochemistry* **58**, 2584–2588 (2019).
42. Müller, I. B. *et al.* The assembly of the plasmodial PLP synthase complex follows a defined course. *PLoS ONE* **3**, e1815 (2008).
43. Dodson, G. & Wlodawer, A. Catalytic triads and their relatives. *Trends Biochem. Sci.* **23**, 347–352 (1998).
44. 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**, E653–E661 (2013).
45. Kneller, D. W. *et al.* Structural plasticity of SARS-CoV-2 3CL Mpro active site cavity revealed by room temperature X-ray crystallography. *Nat. Commun.* **11**, 3202 (2020).
46. Fernandes, H. S., Sousa, S. F. & Cerqueira, N. M. F. S. A. New insights into the catalytic mechanism of the SARS-CoV-2 main protease: an ONIOM QM/MM approach. *Mol. Divers.* **26**, 1373–1381 (2022).
47. Menard, R. *et al.* Importance of hydrogen-bonding interactions involving the side chain of Asp158 in the catalytic mechanism of papain. *Biochemistry* **30**, 5531–5538 (1991).
48. Smith, N. & Wilson, M. A. Understanding Cysteine Chemistry Using Conventional and Serial X-ray Protein Crystallography. *Crystals* **12**, 1671 (2022).

49. Polgár, L. in *Handb. Proteolytic Enzym.* **3**, 2524–2534 (Elsevier, 2013).
50. Gisdon, F. J., Bombarda, E. & Ullmann, G. M. Serine and Cysteine Peptidases: So Similar, Yet Different. How the Active-Site Electrostatics Facilitates Different Reaction Mechanisms. *J. Phys. Chem. B* **126**, 4035–4048 (2022).
51. Hofer, F., Kraml, J., Kahler, U., Kamenik, A. S. & Liedl, K. R. Catalytic Site pKa Values of Aspartic, Cysteine, and Serine Proteases: Constant pH MD Simulations. *J. Chem. Inf. Model.* **60**, 3030–3042 (2020).
52. Kneller, D. W. *et al.* Direct Observation of Protonation State Modulation in SARS-CoV-2 Main Protease upon Inhibitor Binding with Neutron Crystallography. *J. Med. Chem.* **64**, 4991–5000 (2021).
53. Lewis, S. D., Johnson, F. A. & Shafer, J. A. Potentiometric determination of ionizations at the active site of papain. *Biochemistry* **15**, 5009–5017 (1976).
54. Polgár, L. in *Handb. Proteolytic Enzym.* (eds. Rawlings, N. D. & Salvesen, G.) **2**, 1773–1784 (Academic Press, 2013).
55. Ehrenshaft, M., Bilski, P., Li, M. Y., Chignell, C. F. & Daub, M. E. A highly conserved sequence is a novel gene involved in de novo vitamin B6 biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 9374–8 (1999).
56. Osmani, A. H., May, G. S. & Osmani, S. A. The Extremely Conserved pyroA Gene of *Aspergillus nidulans* Is Required for Pyridoxine Synthesis and Is Required Indirectly for Resistance to Photosensitizers. *J. Biol. Chem.* **274**, 23565–23569 (1999).
57. Sakai, A., Kita, M., Katsuragi, T., Ogasawara, N. & Tani, Y. yaaD and yaaE are involved in vitamin B6 biosynthesis in *Bacillus subtilis*. *J. Biosci. Bioeng.* **93**, 309–312 (2002).
58. Robinson, G. C., Kaufmann, M., Roux, C. & Fitzpatrick, T. B. Structural definition of the lysine swing in *Arabidopsis thaliana* PDX1: Intermediate channeling facilitating vitamin B6 biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E5821–E5829 (2016).
59. Rodrigues, M. J. *et al.* Lysine relay mechanism coordinates intermediate transfer in vitamin B6 biosynthesis. *Nat. Chem. Biol.* **13**, 290–294 (2017).
60. Yang, Y., Zhao, G. & Winkler, M. E. Identification of the *pdxK* gene that encodes pyridoxine (vitamin B₆) kinase in *Escherichia coli* K-12. *FEMS Microbiol. Lett.* **141**, 89–95 (1996).
61. Chang, A. *et al.* BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res.* **49**, D498–D508 (2021).
62. Ikushiro, H., Fujii, S., Shiraiwa, Y. & Hayashi, H. Acceleration of the Substrate C α Deprotonation by an Analogue of the Second Substrate Palmitoyl-CoA in Serine Palmitoyltransferase *. *J. Biol. Chem.* **283**, 7542–7553 (2008).
63. Lerma-Ortiz, C. *et al.* ‘Nothing of chemistry disappears in biology’: the Top 30 damage-prone endogenous metabolites. *Biochem. Soc. Trans.* **44**, 961–971 (2016).
64. Mizushina, Y. *et al.* Pyridoxal 5'-phosphate is a selective inhibitor in vivo of DNA polymerase alpha and epsilon. *Biochem. Biophys. Res. Commun.* **312**, 1025–1032 (2003).

65. Vermeersch, J. J. *et al.* Pyridoxal 5'-phosphate inactivates DNA topoisomerase IB by modifying the lysine general acid. *Nucleic Acids Res.* **32**, 5649–5657 (2004).
66. Tremiño, L., Llop, A., Rubio, V. & Contreras, A. The Conserved Family of the Pyridoxal Phosphate-Binding Protein (PLPBP) and Its Cyanobacterial Paradigm PipY. *Life* **12**, 1622 (2022).
67. Hartl, J., Kiefer, P., Meyer, F. & Vorholt, J. A. Longevity of major coenzymes allows minimal de novo synthesis in microorganisms. *Nat. Microbiol.* **2**, 17073 (2017).
68. Liang, J., Han, Q., Tan, Y., Ding, H. & Li, J. Current Advances on Structure-Function Relationships of Pyridoxal 5'-Phosphate-Dependent Enzymes. *Front. Mol. Biosci.* **6**, 4 (2019).
69. Dunathan, H. C. Conformation and reaction specificity in pyridoxal phosphate enzymes. *Proc. Natl. Acad. Sci.* **55**, 712–716 (1966).
70. Dunathan, H. C. in *Adv. Enzymol. Relat. Areas Mol. Biol. Vol. 35* (ed. Meister, A.) **35**, 79–134 (John Wiley & Sons, Inc., 1971).
71. Ngo, P.-T. H. *et al.* Expression, crystallization and preliminary X-ray crystallographic analysis of XometC, a cystathionine γ -lyase-like protein from *Xanthomonas oryzae* pv. *oryzae*. *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.* **64**, 750–753 (2008).
72. Ngo, H. P. T. *et al.* PLP undergoes conformational changes during the course of an enzymatic reaction. *Acta Crystallogr. D Biol. Crystallogr.* **70**, 596–606 (2014).
73. Toney, M. D. Controlling reaction specificity in pyridoxal phosphate enzymes. *Biochim. Biophys. Acta* **1814**, 1407–1418 (2011).
74. Oliveira, E. F., Cerqueira, N. M. F. S. A., Fernandes, P. A. & Ramos, M. J. Mechanism of Formation of the Internal Aldimine in Pyridoxal 5'-Phosphate-Dependent Enzymes. *J. Am. Chem. Soc.* **133**, 15496–15505 (2011).
75. Cerqueira, N. M. F. S. A., Fernandes, P. A. & Ramos, M. J. Computational Mechanistic Studies Addressed to the Transamination Reaction Present in All Pyridoxal 5'-Phosphate-Requiring Enzymes. *J. Chem. Theory Comput.* **7**, 1356–1368 (2011).
76. Li, Z., Zhao, Y., Zhou, H., Luo, H.-B. & Zhan, C.-G. Catalytic Roles of Coenzyme Pyridoxal-5'-phosphate (PLP) in PLP-Dependent Enzymes: Reaction Pathway for Methionine- γ -Lyase-Catalyzed L-Methionine Depletion. *ACS Catal.* **10**, 2198–2210 (2020).
77. Dutta Banik, S., Bankura, A. & Chandra, A. A QM/MM simulation study of transamination reaction at the active site of aspartate aminotransferase: Free energy landscape and proton transfer pathways. *J. Comput. Chem.* **41**, 2684–2694 (2020).
78. Soniya, K. & Chandra, A. Free energy landscapes of prototropic tautomerism in pyridoxal 5'-phosphate schiff bases at the active site of an enzyme in aqueous medium. *J. Comput. Chem.* **39**, 1629–1638 (2018).
79. Soniya, K., Awasthi, S., Nair, N. N. & Chandra, A. Transamination Reaction at the Active Site of Aspartate Aminotransferase: A Proton Hopping Mechanism through Pyridoxal 5'-Phosphate. *ACS Catal.* **9**, 6276–6283 (2019).

80. Rye, C. *et al.* in *Biology* (OpenStax, 2016). at <<https://openstax.org/books/biology/pages/3-5-nucleic-acids>>
81. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).
82. Franklin, R. E. & Gosling, R. G. Molecular Configuration in Sodium Thymonucleate. *Nature* **171**, 740–741 (1953).
83. Franklin, R. E. & Gosling, R. G. The structure of sodium thymonucleate fibres. I. The influence of water content. *Acta Crystallogr.* **6**, 673–677 (1953).
84. Mitsui, Y. *et al.* Physical and Enzymatic Studies on Poly d(I–C).Poly d(I–C), an Unusual Double-helical DNA. *Nature* **228**, 1166–1169 (1970).
85. Guiblet, W. M. *et al.* Non-B DNA: a major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome. *Nucleic Acids Res.* **49**, 1497–1516 (2021).
86. Wang, G. & Vasquez, K. M. Non-B DNA structure-induced genetic instability. *Mutat. Res.* **598**, 103–119 (2006).
87. Wells, R. D. Non-B DNA conformations, mutagenesis and disease. *Trends Biochem. Sci.* **32**, 271–278 (2007).
88. Bacolla, A. & Wells, R. D. Non-B DNA Conformations, Genomic Rearrangements, and Human Disease. *J. Biol. Chem.* **279**, 47411–47414 (2004).
89. Pereira, F. *et al.* Evidence for Variable Selective Pressures at a Large Secondary Structure of the Human Mitochondrial DNA Control Region. *Mol. Biol. Evol.* **25**, 2759–2770 (2008).
90. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465 (1981).
91. Alfonso-Sánchez, M. A. *et al.* Sequence polymorphisms of the mtDNA control region in a human isolate: the Georgians from Swanetia. *J. Hum. Genet.* **51**, 429–439 (2006).
92. Shakhssalim, N. *et al.* The mitochondrial C16069T polymorphism, not mitochondrial D310 (D-loop) mononucleotide sequence variations, is associated with bladder cancer. *Cancer Cell Int.* **13**, 120 (2013).
93. Mousavizadeh, K., Rajabi, P., Alaei, M., Dadgar, S. & Houshmand, M. Usage of mitochondrial D-loop variation to predict risk for Huntington disease. *Mitochondrial DNA* **26**, 579–582 (2015).
94. Jamali, L., Banoei, M. M., Khalili, E., Dadgar, S. & Houshmand, M. Association of genetic variations in the mitochondrial D-loop with β -thalassemia. *Mitochondrial DNA* **27**, 1693–1696 (2016).
95. Damas, J. *et al.* Mitochondrial DNA deletions are associated with non-B DNA conformations. *Nucleic Acids Res.* **40**, 7606–7621 (2012).
96. Gao, F. *et al.* Recent advances in aptamer-based targeted drug delivery systems for cancer therapy. *Front. Bioeng. Biotechnol.* **10**, (2022).

97. Li, Y. *et al.* Aptamer nucleotide analog drug conjugates in the targeting therapy of cancers. *Front. Cell Dev. Biol.* **10**, (2022).
98. Liang, Z. *et al.* The Aptamer Ob2, a novel AChE inhibitor, restores cognitive deficits and alleviates amyloidogenesis in 5×FAD transgenic mice. *Mol. Ther. Nucleic Acids* **28**, 114–123 (2022).
99. Liu, S., Li, S., Lin, J., Li, J. & Yang, H. Aptamer-Induced-Dimerization Strategy Attenuates AβO Toxicity through Modulating the Trophic Activity of PrPC Signaling. *J. Am. Chem. Soc.* **144**, 9264–9270 (2022).
100. Wei, Z., Zhou, Y., Wang, R., Wang, J. & Chen, Z. Aptamers as Smart Ligands for Targeted Drug Delivery in Cancer Therapy. *Pharmaceutics* **14**, 2561 (2022).
101. Zheng, Y. *et al.* Novel DNA Aptamers for Parkinson's Disease Treatment Inhibit α-Synuclein Aggregation and Facilitate its Degradation. *Mol. Ther. Nucleic Acids* **11**, 228–242 (2018).
102. Ray, P. & White, R. R. Aptamers for Targeted Drug Delivery. *Pharmaceutics* **3**, 1761–1778 (2010).
103. Jani, M. S., Veetil, A. T. & Krishnan, Y. Precision immunomodulation with synthetic nucleic acid technologies. *Nat. Rev. Mater.* **4**, 451–458 (2019).
104. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).
105. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505–510 (1990).
106. Stoltenburg, R., Reinemann, C. & Strehlitz, B. SELEX--a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol. Eng.* **24**, 381–403 (2007).
107. Ahmad, N. A., Mohamed Zulkifli, R., Hussin, H. & Nadri, M. H. In silico approach for Post-SELEX DNA aptamers: A mini-review. *J. Mol. Graph. Model.* **105**, 107872 (2021).
108. Buglak, A. A., Samokhvalov, A. V., Zherdev, A. V. & Dzantiev, B. B. Methods and Applications of In Silico Aptamer Design and Modeling. *Int. J. Mol. Sci.* **21**, 8420 (2020).
109. Ni, S. *et al.* Chemical Modifications of Nucleic Acid Aptamers for Therapeutic Purposes. *Int. J. Mol. Sci.* **18**, 1683 (2017).
110. Case, D. A. *et al.* Amber 2021. (2021).
111. Lindahl, Abraham, Hess & van der Spoel. GROMACS 2019.1 Source code. (2019).
112. Phillips, J. C. *et al.* Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **153**, 044130 (2020).
113. Thompson, A. P. *et al.* LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271**, 108171 (2022).
114. Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*. (Wiley, 2004).

115. Atkins, P. W. & Friedman, R. *Molecular Quantum Mechanics*. (Oxford University Press, 2011).
116. Jensen, F. *Introduction to Computational Chemistry*. (John Wiley & Sons, 2017).
117. Levine, I. N. *Quantum chemistry*. (Pearson, 2014).
118. Lewars, E. G. *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*. (Springer International Publishing : Imprint: Springer, 2016). doi:10.1007/978-3-319-30916-3
119. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
120. 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 pKa Values. *J. Chem. Theory Comput.* **7**, 2284–2295 (2011).
121. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **7**, 525–537 (2011).
122. 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).
123. Chen, S. *et al.* A ReaxFF-based molecular dynamics study of the pyrolysis mechanism of hexamethyldisiloxane. *J. Mol. Liq.* **356**, 119026 (2022).
124. Hur, J., Abousleiman, Y. N., Hull, K. L. & Abdolhosseini Qomi, M. J. Reactive force fields for modeling oxidative degradation of organic matter in geological formations. *RSC Adv.* **11**, 29298–29307 (2021).
125. Senftle, T. P. *et al.* The ReaxFF reactive force-field: development, applications and future directions. *Npj Comput. Mater.* **2**, 1–14 (2016).
126. Cornell, W. D. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).
127. Tian, C. *et al.* ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **16**, 528–552 (2020).
128. 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).
129. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
130. Zgarbová, M. *et al.* Refinement of the Sugar–Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *J. Chem. Theory Comput.* **11**, 5723–5736 (2015).
131. Zgarbová, M. *et al.* Refinement of the Cornell *et al.* Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.* **7**, 2886–2902 (2011).

132. 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).
133. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Erratum: 'Development and testing of a general amber force field,' Journal of Computational Chemistry (2004) 25(9) 1157–1174. *J. Comput. Chem.* **26**, 114–114 (2005).
134. *Biomolecular Simulations: Methods and Protocols*. **2022**, (Springer New York, 2019).
135. *Biomolecular Simulations: Methods and Protocols*. **924**, (Humana Press, 2013).
136. Braun, E. *et al.* Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living J. Comput. Mol. Sci.* **1**, (2019).
137. Verlet, L. Computer 'Experiments' on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **159**, 98–103 (1967).
138. Swope, W. C., Andersen, H. C., Berens, P. H. & Wilson, K. R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **76**, 637–649 (1982).
139. Ryckaert, J.-P., 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.* **23**, 327–341 (1977).
140. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).
141. Miyamoto, S. & Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **13**, 952–962 (1992).
142. Hopkins, C. W., Le Grand, S., Walker, R. C. & Roitberg, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theory Comput.* **11**, 1864–1874 (2015).
143. Langevin, P. Sur la théorie du mouvement brownien. *Comptes Rendus Académie Sci.* **146**, 530–533 (1908).
144. 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).
145. Schneider, T. & Stoll, E. Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions. *Phys. Rev. B* **17**, 1302–1322 (1978).
146. Berendsen, H. J. C., Postma, J. P. M., Gunsteren, W. F. van, DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).
147. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).

148. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **61**, 3891–3898 (2021).
149. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* NA-NA (2009). doi:10.1002/jcc.21334
150. Zhang, N. & Zhao, H. Enriching screening libraries with bioactive fragment space. *Bioorg. Med. Chem. Lett.* **26**, 3594–3597 (2016).
151. 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.* **267**, 727–748 (1997).
152. Molecular Operating Software (MOE). (2022).
153. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004).
154. Halgren, T. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **47**, 1750–1759 (2004).
155. Friesner, R. A. *et al.* Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein–Ligand Complexes. *J. Med. Chem.* **49**, 6177–6196 (2006).
156. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A Protein–Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).
157. Honorato, R. V. *et al.* Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem. *Front. Mol. Biosci.* **8**, 729513 (2021).
158. van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **428**, 720–725 (2016).
159. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **10**, 449–461 (2015).
160. Meng, E. C., Shoichet, B. K. & Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **13**, 505–524 (1992).
161. Ewing, T. J. A. & Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **18**, 1175–1189 (1997).
162. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V. & Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* **11**, 425–445 (1997).
163. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins Struct. Funct. Bioinforma.* **52**, 609–623 (2003).

164. Gehlhaar, D. K. *et al.* Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **2**, 317–324 (1995).
165. Verkhivker, G. M. *et al.* Monte Carlo simulations of the peptide recognition at the consensus binding site of the constant fragment of human immunoglobulin G: The energy landscape analysis of a hot spot at the intermolecular interface. *Proteins Struct. Funct. Genet.* **48**, 539–557 (2002).
166. Verkhivker, G. M. *et al.* Computational detection of the binding-site hot spot at the remodeled human growth hormone-receptor interface. *Proteins Struct. Funct. Genet.* **53**, 201–219 (2003).
167. Verkhivker, G. M. Computational analysis of ligand binding dynamics at the intermolecular hot spots with the aid of simulated tempering and binding free energy calculations. *J. Mol. Graph. Model.* **22**, 335–348 (2004).
168. Mooij, W. T. M. & Verdonk, M. L. General and targeted statistical potentials for protein-ligand interactions. *Proteins Struct. Funct. Bioinforma.* **61**, 272–287 (2005).
169. Kalodimos, C. G. *et al.* Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes. *Science* **305**, 386–389 (2004).
170. Kopke Salinas, R. *et al.* Altered Specificity in DNA Binding by the lac Repressor: A Mutant lac Headpiece that Mimics the gal Repressor. *ChemBioChem* **6**, 1628–1637 (2005).
171. Kamphuis, M. B. *et al.* Model for RNA Binding and the Catalytic Site of the RNase Kid of the Bacterial parD Toxin–Antitoxin System. *J. Mol. Biol.* **357**, 115–126 (2006).
172. van Dijk, M., van Dijk, A. D. J., Hsu, V., Boelens, R. & Bonvin, A. M. J. J. Information-driven protein–DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Res.* **34**, 3317–3325 (2006).
173. van Dijk, M. & Bonvin, A. M. J. J. A protein–DNA docking benchmark. *Nucleic Acids Res.* **36**, e88 (2008).
174. Repanas, K., Fuentes, G., Cohen, S. X., Bonvin, A. M. J. J. & Perrakis, A. Insights into the DNA cleavage mechanism of human LINE-1 retrotransposon endonuclease. *Proteins Struct. Funct. Bioinforma.* **74**, 917–928 (2009).
175. Romanuka, J. *et al.* Specificity and Affinity of Lac Repressor for the Auxiliary Operators O2 and O3 Are Explained by the Structures of Their Protein–DNA Complexes. *J. Mol. Biol.* **390**, 478–489 (2009).
176. Kobayashi, M., Ab, E., Bonvin, A. M. J. J. & Siegal, G. Structure of the DNA-bound BRCA1 C-terminal Region from Human Replication Factor C p140 and Model of the Protein-DNA Complex. *J. Biol. Chem.* **285**, 10087–10097 (2010).
177. van Dijk, M. & Bonvin, A. M. J. J. Pushing the limits of what is achievable in protein–DNA docking: benchmarking HADDOCK’s performance. *Nucleic Acids Res.* **38**, 5634–5647 (2010).
178. van Dijk, M., Visscher, K. M., Kastitis, P. L. & Bonvin, A. M. J. J. Solvated protein–DNA docking using HADDOCK. *J. Biomol. NMR* **56**, 51–63 (2013).

179. Wu, A. M. *et al.* Activity–structure correlations in divergent lectin evolution: fine specificity of chicken galectin CG-14 and computational analysis of flexible ligand docking for CG-14 and the closely related CG-16. *Glycobiology* **17**, 165–184 (2007).
180. Krzeminski, M. *et al.* Human galectin-3 (Mac-2 antigen): Defining molecular switches of affinity to natural glycoproteins, structural and dynamic aspects of glycan binding by flexible ligand docking and putative regulatory sequences in the proximal promoter region. *Biochim. Biophys. Acta BBA - Gen. Subj.* **1810**, 150–161 (2011).
181. Dagil, R., O'Shea, C., Nykjær, A., Bonvin, A. M. J. J. & Kragelund, B. B. Gentamicin Binds to the Megalin Receptor as a Competitive Inhibitor Using the Common Ligand Binding Motif of Complement Type Repeats: INSIGHT FROM THE NMR STRUCTURE OF THE 10TH COMPLEMENT TYPE REPEAT DOMAIN ALONE AND IN COMPLEX WITH GENTAMICIN *. *J. Biol. Chem.* **288**, 4424–4435 (2013).
182. Kurkcuoglu, Z. *et al.* Performance of HADDOCK and a simple contact-based protein–ligand binding affinity predictor in the D3R Grand Challenge 2. *J. Comput. Aided Mol. Des.* **32**, 175–185 (2018).
183. Koukos, P. I., Xue, L. C. & Bonvin, A. M. J. J. Protein–ligand pose and affinity prediction: Lessons from D3R Grand Challenge 3. *J. Comput. Aided Mol. Des.* **33**, 83–91 (2019).
184. Vangone, A. *et al.* Large-scale prediction of binding affinity in protein–small ligand complexes: the PRODIGY-LIG web server. *Bioinformatics* **35**, 1585–1587 (2019).
185. Vangone, A. *et al.* Sense and simplicity in HADDOCK scoring: Lessons from CASP-CAPRI round 1. *Proteins Struct. Funct. Bioinforma.* **85**, 417–423 (2017).
186. Vangone, A. *et al.* Sense and Simplicity in HADDOCK Scoring: Lessons from CASP-CAPRI (page 418). *Proteins Struct. Funct. Bioinforma.* **85**, 1589–1590 (2017).
187. Nguyen, H., Pérez, A., Bermeo, S. & Simmerling, C. Refinement of Generalized Born Implicit Solvation Parameters for Nucleic Acids and Their Complexes with Proteins. *J. Chem. Theory Comput.* **11**, 3714–3728 (2015).
188. Mongan, J., Simmerling, C., McCammon, J. A., Case, D. A. & Onufriev, A. Generalized Born Model with a Simple, Robust Molecular Volume Correction. *J. Chem. Theory Comput.* **3**, 156–169 (2007).
189. Roe, D. R., Okur, A., Wickstrom, L., Hornak, V. & Simmerling, C. Secondary Structure Bias in Generalized Born Solvent Models: Comparison of Conformational Ensembles and Free Energy of Solvent Polarization from Explicit and Implicit Solvation. *J. Phys. Chem. B* **111**, 1846–1857 (2007).
190. Weiser, J., Shenkin, P. S. & Still, W. C. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J. Comput. Chem.* **20**, 217–230 (1999).
191. Planck, M. Ueber die Elementarquanten der Materie und der Elektrizität. *Ann. Phys.* **309**, 564–566 (1901).
192. Einstein, A. Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Ann. Phys.* **322**, 132–148 (1905).
193. Heisenberg, W. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Z. Für Phys.* **43**, 172–198 (1927).

194. Schrödinger, E. Quantisierung als Eigenwertproblem. *Ann. Phys.* **384**, 361–376 (1926).
195. Born, M. Zur Quantenmechanik der Stoßvorgänge. *Z. Für Phys.* **37**, 863–867 (1926).
196. Born, M. & Oppenheimer, R. Zur Quantentheorie der Molekeln. *Ann. Phys.* **389**, 457–484 (1927).
197. Ditchfield, R., Hehre, W. J. & Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **54**, 724–728 (1971).
198. Hehre, W. J., Ditchfield, R. & Pople, J. A. Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **56**, 2257–2261 (1972).
199. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297 (2005).
200. Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **90**, 1007–1023 (1989).
201. Prejanò, M., Marino, T. & Russo, N. QM Cluster or QM/MM in Computational Enzymology: The Test Case of LigW-Decarboxylase. *Front. Chem.* **6**, (2018).
202. Fernandes, H. S., Ramos, M. J. & Cerqueira, N. M. F. S. A. Catalytic Mechanism of the Serine Hydroxymethyltransferase: A Computational ONIOM QM/MM Study. *ACS Catal.* **8**, 10096–10110 (2018).
203. Silva, C. F., Borges, K. B. & Nascimento, C. S. Computational study on acetamiprid-molecular imprinted polymer. *J. Mol. Model.* **25**, 104 (2019).
204. Teixeira, C., Ramos, M., Sousa, S. & Cerqueira, N. Solving the Catalytic Mechanism of Tryptophan Synthase: an Emergent Drug Target in the Treatment of Tuberculosis. *ChemCatChem* **12**, 227–237 (2020).
205. Teixeira, C. S. S., Sousa, S. F. & Cerqueira, N. M. F. S. A. An Unusual Cys-Glu-Lys Catalytic Triad is Responsible for the Catalytic Mechanism of the Nitrilase Superfamily: A QM/MM Study on Nit2. *ChemPhysChem* **22**, 796–804 (2021).
206. Ribeiro, P. *et al.* The complete catalytic mechanism of xanthine oxidase: a computational study. *Inorg. Chem. Front.* **8**, 405–416 (2021).
207. Pina, A. F., Sousa, S. F. & Cerqueira, N. M. F. S. A. The Catalytic Mechanism of Pdx2 Glutaminase Driven by a Cys-His-Glu Triad: A Computational Study. *ChemBioChem* **23**, e202100555 (2021).
208. Rocha, J. F., Sousa, S. F. & Cerqueira, N. M. F. S. A. Computational Studies Devoted to the Catalytic Mechanism of Threonine Aldolase, a Critical Enzyme in the Pharmaceutical Industry to Synthesize β -Hydroxy- α -amino Acids. *ACS Catal.* **12**, 4990–4999 (2022).
209. Almatarneh, M. H., Kayed, G. G., Altarawneh, M., Zhao, Y. & Verma, A. Computational Insights in DNA Methylation: Catalytic and Mechanistic Elucidations for Forming 3-Methyl Cytosine. *J. Chem.* **2022**, e2673396 (2022).

210. Magalhães, R. P., Fernandes, H. S. & Sousa, S. F. The critical role of Asp206 stabilizing residues on the catalytic mechanism of the *Ideonella sakaiensis* PETase. *Catal. Sci. Technol.* **12**, 3474–3483 (2022).
211. Hariharan, P. C. & Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor. Chim. Acta* **28**, 213–222 (1973).
212. Clark, T., Chandrasekhar, J., Spitznagel, G. W. & Schleyer, P. V. R. Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements, Li–F. *J. Comput. Chem.* **4**, 294–301 (1983).
213. Kendall, R. A., Dunning, T. H. & Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **96**, 6796–6806 (1992).
214. Hartree, D. R. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part II. Some Results and Discussion. *Math. Proc. Camb. Philos. Soc.* **24**, 111–132 (1928).
215. Pauli, W. Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren. *Z. Für Phys.* **31**, 765–783 (1925).
216. Slater, J. C. The Theory of Complex Spectra. *Phys. Rev.* **34**, 1293–1322 (1929).
217. Møller, Chr. & Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **46**, 618–622 (1934).
218. Čížek, J. On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods. *J. Chem. Phys.* **45**, 4256–4266 (1966).
219. Riplinger, C. & Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.* **138**, 034106 (2013).
220. Riplinger, C., Sandhoefer, B., Hansen, A. & Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *J. Chem. Phys.* **139**, 134101 (2013).
221. 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).
222. Liakos, D. G., Sparta, M., Kesharwani, M. K., Martin, J. M. L. & Neese, F. Exploring the Accuracy Limits of Local Pair Natural Orbital Coupled-Cluster Theory. *J. Chem. Theory Comput.* **11**, 1525–1539 (2015).
223. 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**, 024109 (2016).
224. Hohenberg, P. & Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **136**, B864–B871 (1964).
225. Kohn, W. & Sham, L. J. Quantum Density Oscillations in an Inhomogeneous Electron Gas. *Phys. Rev.* **137**, A1697–A1705 (1965).

226. Kohn, W. & Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
227. Perdew, J. P. & Schmidt, K. Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conf. Proc.* **577**, 1–20 (2001).
228. Bursch, M., Mewes, J., Hansen, A. & Grimme, S. Best-Practice DFT Protocols for Basic Molecular Computational Chemistry. *Angew. Chem. Int. Ed.* (2022). doi:10.1002/anie.202205735
229. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
230. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Phys. Rev. Lett.* **78**, 1396–1396 (1997).
231. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38**, 3098–3100 (1988).
232. 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).
233. Tao, J., Perdew, J. P., Staroverov, V. N. & Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta-Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **91**, 146401 (2003).
234. Zhao, Y. & Truhlar, D. G. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.* **125**, 194101 (2006).
235. Peverati, R. & Truhlar, D. G. M11-L: A Local Density Functional That Provides Improved Accuracy for Electronic Structure Calculations in Chemistry and Physics. *J. Phys. Chem. Lett.* **3**, 117–124 (2012).
236. Peverati, R. & Truhlar, D. G. An improved and broadly accurate local approximation to the exchange–correlation density functional: The MN12-L functional for electronic structure calculations in chemistry and physics. *Phys. Chem. Chem. Phys.* **14**, 13171–13174 (2012).
237. Furness, J. W., Kaplan, A. D., Ning, J., Perdew, J. P. & Sun, J. Accurate and Numerically Efficient r2SCAN Meta-Generalized Gradient Approximation. *J. Phys. Chem. Lett.* **11**, 8208–8215 (2020).
238. Furness, J. W., Kaplan, A. D., Ning, J., Perdew, J. P. & Sun, J. Correction to “Accurate and Numerically Efficient r2SCAN Meta-Generalized Gradient Approximation”. *J. Phys. Chem. Lett.* **11**, 9248–9248 (2020).
239. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98**, 5648–5652 (1993).
240. Stephens, P. J., Devlin, F. J., Chabalowski, C. F. & Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **98**, 11623–11627 (1994).
241. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132**, 154104 (2010).

242. Caldeweyher, E., Bannwarth, C. & Grimme, S. Extension of the D3 dispersion coefficient model. *J. Chem. Phys.* **147**, 034112 (2017).
243. Kruse, H. & Grimme, S. A geometrical correction for the inter- and intra-molecular basis set superposition error in Hartree-Fock and density functional theory calculations for large systems. *J. Chem. Phys.* **136**, 154101 (2012).
244. Brandenburg, J. G. *et al.* Geometrical Correction for the Inter- and Intramolecular Basis Set Superposition Error in Periodic Density Functional Theory Calculations. *J. Phys. Chem. A* **117**, 9282–9292 (2013).
245. Grimme, S., Brandenburg, J. G., Bannwarth, C. & Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *J. Chem. Phys.* **143**, 054107 (2015).
246. Brandenburg, J. G., Bannwarth, C., Hansen, A. & Grimme, S. B97-3c: A revised low-cost variant of the B97-D density functional method. *J. Chem. Phys.* **148**, 064104 (2018).
247. Grimme, S., Hansen, A., Ehlert, S. & Mewes, J.-M. r²SCAN-3c: A “Swiss army knife” composite electronic-structure method. *J. Chem. Phys.* **154**, 064103 (2021).
248. Vydrov, O. A. & Van Voorhis, T. Nonlocal van der Waals density functional: The simpler the better. *J. Chem. Phys.* **133**, 244103 (2010).
249. Dewar, M. J. S., Zoebisch, E. G., Healy, E. F. & Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **107**, 3902–3909 (1985).
250. Stewart, J. J. P. Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem.* **10**, 209–220 (1989).
251. Stewart, J. J. P. Optimization of parameters for semiempirical methods II. Applications. *J. Comput. Chem.* **10**, 221–264 (1989).
252. Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **13**, 1173–1213 (2007).
253. Gaus, M., Cui, Q. & Elstner, M. DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *J. Chem. Theory Comput.* **7**, 931–948 (2011).
254. Grimme, S., Bannwarth, C. & Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86). *J. Chem. Theory Comput.* **13**, 1989–2009 (2017).
255. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
256. Schlegel, H. B. in *Mod. Electron. Struct. Theory* **2**, 459–500 (World Scientific Publishing Company, 1995).

257. Fernandes, H. S., Ramos, M. J. & M. F. S. A. Cerqueira, N. molUP: A VMD plugin to handle QM and ONIOM calculations using the gaussian software. *J. Comput. Chem.* **39**, 1344–1353 (2018).
258. 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).
259. Ochterski, J. W. Thermochemistry in Gaussian. *Gaussian Inc* 1–19 (2000).
260. 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).
261. Reuter, N., Dejaegere, A., Maignet, B. & Karplus, M. Frontier Bonds in QM/MM Methods: A Comparison of Different Approaches. *J. Phys. Chem. A* **104**, 1720–1735 (2000).
262. Cui, Q., Elstner, M., Kaxiras, E., Frauenheim, T. & Karplus, M. A QM/MM Implementation of the Self-Consistent Charge Density Functional Tight Binding (SCC-DFTB) Method. *J. Phys. Chem. B* **105**, 569–585 (2001).
263. The Nobel Prize in Chemistry 2013. *NobelPrize.org* at <<https://www.nobelprize.org/prizes/chemistry/2013/summary/>>
264. 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**, 1170–1179 (1995).
265. Arrhenius, S. Über die Dissociationswärme und den Einfluss der Temperatur auf den Dissociationsgrad der Elektrolyte. *Z. Für Phys. Chem.* **4U**, 96–116 (1889).
266. Eyring, H. The Activated Complex in Chemical Reactions. *J. Chem. Phys.* **3**, 107–115 (1935).
267. World Health Organization. *World malaria report 2020: 20 years of global progress and challenges*. (World Health Organization, 2020).
268. World Health Organization. *Global tuberculosis report 2020*. (World Health Organization, 2020).
269. Wrenger, C., Knöckel, J., Walter, R. D. & Müller, I. B. Vitamin B1 and B6 in the malaria parasite: Requisite or dispensable? *Braz. J. Med. Biol. Res.* **41**, 82–88 (2008).
270. Barra, A. L. C. *et al.* Essential Metabolic Routes as a Way to ESKAPE From Antibiotic Resistance. *Front. Public Health* **8**, 26 (2020).
271. Kumar, P. *et al.* L-Threonine Transaldolase Activity Is Enabled by a Persistent Catalytic Intermediate. *ACS Chem. Biol.* **16**, 86–95 (2021).
272. Rubinstein, A. & Major, D. T. Understanding Catalytic Specificity in Alanine Racemase from Quantum Mechanical and Molecular Mechanical Simulations of the Arginine 219 Mutant. *Biochemistry* **49**, 3957–3964 (2010).

273. Hill, R. E. *et al.* The Biogenetic Anatomy of Vitamin B6: A ^{13}C NMR INVESTIGATION OF THE BIOSYNTHESIS OF PYRIDOXOL IN *ESCHERICHIA COLI**. *J. Biol. Chem.* **271**, 30426–30435 (1996).
274. Kennedy, I. A., Hill, R. E., Pauloski, R. M., Sayer, B. G. & Spenser, I. D. Biosynthesis of Vitamin B6: Origin of Pyridoxine by the Union of Two Acyclic Precursors, 1-Deoxy-D-xylulose and 4-Hydroxy-L-threonine. *J. Am. Chem. Soc.* **117**, 1661–1662 (1995).
275. Thoden, J. B. *et al.* Carbamoyl Phosphate Synthetase: Caught in the Act of Glutamine Hydrolysis. *Biochemistry* **37**, 8825–8831 (1998).
276. Thoden, J. B., Huang, X., Raushel, F. M. & Holden, H. M. The small subunit of carbamoyl phosphate synthetase: snapshots along the reaction pathway. *Biochemistry* **38**, 16158–16166 (1999).
277. Roux, B. & Walsh, C. T. p-Aminobenzoate synthesis in *Escherichia coli*: Mutational analysis of three conserved amino acid residues of the amidotransferase PabA. *Biochemistry* **32**, 3763–3768 (1993).
278. Hewagama, A., Guy, H. I., Chaparian, M. & Evans, D. R. The function of Glu338 in the catalytic triad of the carbamoyl phosphate synthetase amidotransferase domain. *Biochim. Biophys. Acta BBA - Protein Struct. Mol. Enzymol.* **1388**, 489–499 (1998).
279. Huang, X. & Raushel, F. M. Deconstruction of the Catalytic Array within the Amidotransferase Subunit of Carbamoyl Phosphate Synthetase. *Biochemistry* **38**, 15909–15914 (1999).
280. Schnizer, H. G., Boehlein, S. K., Stewart, J. D., Richards, N. G. J. & Schuster, S. M. Gamma-Glutamyl thioester intermediate in glutaminase reaction catalyzed by *Escherichia coli* asparagine synthetase B. *Methods Enzymol.* **354**, 260–271 (2002).
281. Hart, E. J. & Powers-Lee, S. G. Mutation analysis of carbamoyl phosphate synthetase: Does the structurally conserved glutamine amidotransferase triad act as a functional dyad? *Protein Sci.* **17**, 1120–1128 (2008).
282. Laidler, K. J. & King, M. C. Development of transition-state theory. *J. Phys. Chem.* **87**, 2657–2664 (1983).
283. Case, D. A. *et al.* Amber 2018. (2018).
284. Frisch, M. J. *et al.* Gaussian 09. (2016).
285. Oda, A., Fukuyoshi, S., Nakagaki, R. & Takahashi, O. Determination of AMBER Force Field Parameters for Thioester by Quantum Chemical Calculations. *Chem. Lett.* **42**, 1206–1208 (2013).
286. 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).
287. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
288. Fernandes, H. S., Sousa, S. F. & Cerqueira, N. M. F. S. A. VMD Store – A VMD Plugin to Browse, Discover, and Install VMD Extensions. *J. Chem. Inf. Model.* **59**, 4519–4523 (2019).

289. Dapprich, S., Komáromi, I., Byun, K. S., Morokuma, K. & Frisch, M. J. A new ONIOM implementation in Gaussian98. Part I. The calculation of energies, gradients, vibrational frequencies and electric field derivatives¹Dedicated to Professor Keiji Morokuma in celebration of his 65th birthday.1. *J. Mol. Struct. THEOCHEM* **461–462**, 1–21 (1999).
290. 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.* **100**, 19357–19363 (1996).
291. Francl, M. M. *et al.* Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *J. Chem. Phys.* **77**, 3654–3665 (1982).
292. Gordon, M. S., Binkley, J. S., Pople, J. A., Pietro, W. J. & Hehre, W. J. Self-consistent molecular-orbital methods. 22. Small split-valence basis sets for second-row elements. *J. Am. Chem. Soc.* **104**, 2797–2803 (1982).
293. Fukui, K. The path of chemical reactions - the IRC approach. *Acc. Chem. Res.* **14**, 363–368 (1981).
294. Hratchian, H. P. & Schlegel, H. B. in *Theory Appl. Comput. Chem.* (eds. Dykstra, C. E., Frenking, G., Kim, K. S. & Scuseria, G. E.) 195–249 (Elsevier, 2005).
295. Neese, F. The ORCA program system. *WIREs Comput. Mol. Sci.* **2**, 73–78 (2012).
296. Neese, F. Software update: the ORCA program system, version 4.0. *WIREs Comput. Mol. Sci.* **8**, e1327 (2018).
297. Woon, D. E. & Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *J. Chem. Phys.* **98**, 1358–1371 (1993).
298. Weigend, F., Köhn, A. & Hättig, C. Efficient use of the correlation consistent basis sets in resolution of the identity MP2 calculations. *J. Chem. Phys.* **116**, 3175–3183 (2002).
299. Sousa, J., Neves, R., Sousa, S., Ramos, M. & Fernandes, P. Reaction Mechanism and Determinants for Efficient Catalysis by DszB, a Key Enzyme for Crude Oil Bio-desulfurization. *ACS Catal.* **10**, 9545–9554 (2020).
300. Limas, N. G. & Manz, T. A. Introducing DDEC6 atomic population analysis: part 2. Computed results for a wide range of periodic and nonperiodic materials. *RSC Adv.* **6**, 45727–45747 (2016).
301. Manz, T. A. & Limas, N. G. Introducing DDEC6 atomic population analysis: part 1. Charge partitioning theory and methodology. *RSC Adv.* **6**, 47771–47801 (2016).
302. Manz, T. A. & Limas, N. G. Chargemol program for performing DDEC analysis. (2017). at <ddec.sourceforge.net>
303. Chaudhuri, B. N. *et al.* Crystal Structure of Imidazole Glycerol Phosphate Synthase. *Structure* **9**, 987–997 (2001).
304. Douangamath, A. *et al.* Structural Evidence for Ammonia Tunneling across the (β α)₈ Barrel of the Imidazole Glycerol Phosphate Synthase Bienenzyme Complex. *Structure* **10**, 185–193 (2002).

305. Spraggon, G. *et al.* The structures of anthranilate synthase of *Serratia marcescens* crystallized in the presence of (i) its substrates, chorismate and glutamine, and a product, glutamate, and (ii) its end-product inhibitor, L -tryptophan. *Proc. Natl. Acad. Sci.* **98**, 6021–6026 (2001).
306. Cerqueira, N. M. F. S. A., Fernandes, P. A., Gonzalez, P. J., Moura, J. J. G. & Ramos, M. J. The Sulfur Shift: An Activation Mechanism for Periplasmic Nitrate Reductase and Formate Dehydrogenase. *Inorg. Chem.* **52**, 10766–10772 (2013).
307. Cerqueira, N. M. F. S. A., Moorthy, H., Fernandes, P. A. & Ramos, M. J. The mechanism of the Ser-(cis)Ser-Lys catalytic triad of peptide amidases. *Phys. Chem. Chem. Phys.* **19**, 12343–12354 (2017).
308. Sundaram, R. M. *et al.* Update on Bacterial Blight of Rice: Fourth International Conference on Bacterial Blight. *Rice* **7**, 12, s12284-014-0012–7 (2014).
309. Kumar, A. *et al.* Deployment of Genetic and Genomic Tools Toward Gaining a Better Understanding of Rice-Xanthomonas oryzae Interactions for Development of Durable Bacterial Blight Resistant Rice. *Front. Plant Sci.* **11**, 1152 (2020).
310. Ito, T. *et al.* Mechanism of eukaryotic serine racemase-catalyzed serine dehydration. *Biochim. Biophys. Acta - Proteins Proteomics* **1868**, 140460 (2020).
311. Qin, Z., Yan, Q., Ma, Q. & Jiang, Z. Crystal structure and characterization of a novel L-serine ammonia-lyase from *Rhizomucor miehei*. *Biochem. Biophys. Res. Commun.* **466**, 431–437 (2015).
312. Yamada, T. *et al.* Crystal Structure of Serine Dehydratase from Rat Liver. *Biochemistry* **42**, 12854–12865 (2003).
313. Sun, L. *et al.* Crystallization and preliminary crystallographic analysis of human serine dehydratase. *Acta Crystallogr. D Biol. Crystallogr.* **59**, 2297–2299 (2003).
314. Wang, C.-Y., Ku, S. C., Lee, C.-C. & Wang, A. H.-J. Modulating the function of human serine racemase and human serine dehydratase by protein engineering. *Protein Eng. Des. Sel.* **25**, 741–749 (2012).
315. Messerschmidt, A. *et al.* Determinants of Enzymatic Specificity in the Cys-Met-Metabolism PLP-Dependent Enzyme Family: Crystal Structure of Cystathionine γ -Lyase from Yeast and Intrafamilial Structure Comparison. *Biol. Chem.* **384**, 373–386 (2003).
316. 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).
317. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **113**, 6378–6396 (2009).
318. Neese, F., Wennmohs, F., Becker, U. & Riplinger, C. The ORCA quantum chemistry program package. *J. Chem. Phys.* **152**, 224108 (2020).
319. Neese, F. Software update: The ORCA program system — Version 5.0. *WIREs Comput. Mol. Sci.* **12**, e1606 (2022).

320. Neese, F. *et al.* ORCA – An ab initio, DFT and semiempirical SCF-MO package, Version 5.0.3. (2022).
321. Case, D. A. *et al.* Amber 2019. (2019).
322. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **97**, 10269–10280 (1993).
323. Cornell, W. D., Cieplak, P., Bayly, C. I. & Kollman, P. A. Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *J. Am. Chem. Soc.* **115**, 9620–9631 (1993).
324. Cieplak, P., Cornell, W. D., Bayly, C. & Kollman, P. A. Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. *J. Comput. Chem.* **16**, 1357–1377 (1995).
325. Horn, A. H. C. A consistent force field parameter set for zwitterionic amino acid residues. *J. Mol. Model.* **20**, 2478 (2014).
326. Machado, M. R. & Pantano, S. Split the Charge Difference in Two! A Rule of Thumb for Adding Proper Amounts of Ions in MD Simulations. *J. Chem. Theory Comput.* **16**, 1367–1372 (2020).
327. Nocedal, J. Updating quasi-Newton matrices with limited storage. *Math. Comput.* **35**, 773–782 (1980).
328. Banerjee, A., Adams, N., Simons, J. & Shepard, R. Search for stationary points on surfaces. *J. Phys. Chem.* **89**, 52–57 (1985).
329. Maurer, L. R., Bursch, M., Grimme, S. & Hansen, A. Assessing Density Functional Theory for Chemically Relevant Open-Shell Transition Metal Reactions. *J. Chem. Theory Comput.* **17**, 6134–6151 (2021).
330. Henkelman, G. & Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **113**, 9978–9985 (2000).
331. Jónsson, H., Mills, G. & Jacobsen, K. W. Nudged elastic band method for finding minimum energy paths of transitions. in *Class. Quantum Dyn. Condens. Phase Simul.* 385–404 (WORLD SCIENTIFIC, 1998). doi:10.1142/9789812839664_0016
332. Mills, G., Jónsson, H. & Schenter, G. K. Reversible work transition state theory: application to dissociative adsorption of hydrogen. *Surf. Sci.* **324**, 305–337 (1995).
333. Ásgeirsson, V. *et al.* Nudged Elastic Band Method for Molecular Reactions Using Energy-Weighted Springs Combined with Eigenvector Following. *J. Chem. Theory Comput.* **17**, 4929–4945 (2021).
334. Bannwarth, C. *et al.* Extended tight-binding quantum chemistry methods. *WIREs Comput. Mol. Sci.* **11**, (2021).
335. Lu, T. & Chen, F. Multiwfn: A multifunctional wavefunction analyzer. *J. Comput. Chem.* **33**, 580–592 (2012).
336. Lu, T. & Chen, F. Atomic Dipole Moment Corrected Hirshfeld Population Method. *J. Theor. Comput. Chem.* **11**, 163–183 (2012).

337. Adamo, C. & Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **110**, 6158–6170 (1999).
338. Ernzerhof, M. & Scuseria, G. E. Assessment of the Perdew–Burke–Ernzerhof exchange–correlation functional. *J. Chem. Phys.* **110**, 5029–5036 (1999).
339. Huang, S. *et al.* Site-Directed Mutagenesis on Human Cystathionine- γ -Lyase Reveals Insights into the Modulation of H₂S Production. *J. Mol. Biol.* **396**, 708–718 (2010).
340. Borchert, A. J., Dustin C. Ernst, Ernst, D. & Downs, D. M. Reactive Enamines and Imines In Vivo: Lessons from the RidA Paradigm. *Trends Biochem. Sci.* **44**, 849–860 (2019).
341. Irons, J. L., Hodge-Hanson, K. & Downs, D. M. RidA Proteins Protect against Metabolic Damage by Reactive Intermediates. *Microbiol. Mol. Biol. Rev.* **84**, e00024-20 (2020).
342. Flavin, M. & Slaughter, C. An Intermediate Trapped by Maleimides in a Pyridoxal-Phosphate Potentiated Enzymatic Elimination Reaction. *Biochemistry* **3**, 885–893 (1964).
343. Darbyshire, A. L., Mothersole, R. G. & Wolthers, K. R. A Fold Type II PLP-Dependent Enzyme from *Fusobacterium nucleatum* Functions as a Serine Synthase and Cysteine Synthase. *Biochemistry* **60**, 524–536 (2021).
344. Huang, Y. *et al.* Pyridoxal-5'-phosphate as an oxygenase cofactor: Discovery of a carboxamide-forming, α -amino acid monooxygenase-decarboxylase. *Proc. Natl. Acad. Sci.* **115**, 974–979 (2018).
345. Petrosino, M. *et al.* H₂S biogenesis by cystathionine beta-synthase: mechanism of inhibition by aminooxyacetic acid and unexpected role of serine. *Cell. Mol. Life Sci.* **79**, 438 (2022).
346. Sagong, H.-Y. & Kim, K.-J. Lysine Decarboxylase with an Enhanced Affinity for Pyridoxal 5-Phosphate by Disulfide Bond-Mediated Spatial Reconstitution. *PLOS ONE* **12**, e0170163 (2017).
347. Wadsworth, J. M. *et al.* The Chemical Basis of Serine Palmitoyltransferase Inhibition by Myriocin. *J. Am. Chem. Soc.* **135**, 14276–14285 (2013).
348. Raman, M. C. C. *et al.* The External Aldimine Form of Serine Palmitoyltransferase. *J. Biol. Chem.* **284**, 17328–17339 (2009).
349. Maenpuen, S. *et al.* Kinetic Mechanism and the Rate-limiting Step of *Plasmodium vivax* Serine Hydroxymethyltransferase. *J. Biol. Chem.* **290**, 8656–8665 (2015).
350. Metzler, C. M., Harris, A. G. & Metzler, D. E. Spectroscopic studies of quinonoid species from pyridoxal 5'-phosphate. *Biochemistry* **27**, 4923–4933 (1988).
351. Lambrecht, J. A., Flynn, J. M. & Downs, D. M. Conserved YjgF protein family deaminates reactive enamine/imine intermediates of pyridoxal 5'-phosphate (PLP)-dependent enzyme reactions. *J. Biol. Chem.* **287**, 3454–3461 (2012).
352. Lambrecht, J. A., Schmitz, G. E. & Downs, D. M. RidA Proteins Prevent Metabolic Damage Inflicted by PLP-Dependent Dehydratases in All Domains of Life. *mBio* **4**, e00033-13 (2013).

353. Liu, X., Zeng, J., Chen, X. & Xie, W. Crystal structures of RidA, an important enzyme for the prevention of toxic side products. *Sci. Rep.* **6**, 30494 (2016).
354. Gao, B. *et al.* Remodeling the conformational dynamics of I-motif DNA by helicases in ATP-independent mode at acidic environment. *iScience* **25**, 103575 (2022).
355. Li, M.-R., Zhang, N. & Zhang, F.-S. Computational investigation of the conformation transitions of DNA in modified water models. *J. Mol. Liq.* **271**, 175–181 (2018).
356. Savelyev, A. & MacKerell, A. D. Differential Impact of the Monovalent Ions Li^+ , Na^+ , K^+ , and Rb^+ on DNA Conformational Properties. *J. Phys. Chem. Lett.* **6**, 212–216 (2015).
357. Shen, X., Atamas, N. A. & Zhang, F. S. Competition between Na^+ and Rb^+ in the minor groove of DNA. *Phys. Rev. E* **85**, 051913 (2012).
358. Singh, A. & Singh, N. Effect of salt concentration on the stability of heterogeneous DNA. *Phys. Stat. Mech. Its Appl.* **419**, 328–334 (2015).
359. Dans, P. D. *et al.* Long-timescale dynamics of the Drew–Dickerson dodecamer. *Nucleic Acids Res.* **44**, 4052–4066 (2016).
360. Hong, F., Schreck, J. S. & Šulc, P. Understanding DNA interactions in crowded environments with a coarse-grained model. *Nucleic Acids Res.* **48**, 10726–10738 (2020).
361. Nakano, S., Miyoshi, D. & Sugimoto, N. Effects of Molecular Crowding on the Structures, Interactions, and Functions of Nucleic Acids. *Chem. Rev.* **114**, 2733–2758 (2014).
362. Singh, A. & Singh, N. DNA melting in the presence of molecular crowders. *Phys. Chem. Chem. Phys.* **19**, 19452–19460 (2017).
363. Miyoshi, D., Nakao, A. & Sugimoto, N. Molecular Crowding Regulates the Structural Switch of the DNA G-Quadruplex. *Biochemistry* **41**, 15017–15024 (2002).
364. Ozcan, Kadir. A., Ghaffari, L. T. & Haeusler, A. R. The effects of molecular crowding and CpG hypermethylation on DNA G-quadruplexes formed by the C9orf72 nucleotide repeat expansion. *Sci. Rep.* **11**, 23213 (2021).
365. Jiang, H.-X. *et al.* Divalent cations and molecular crowding buffers stabilize G-triplex at physiologically relevant temperatures. *Sci. Rep.* **5**, 9255 (2015).
366. Khristich, A. N. & Mirkin, S. M. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *J. Biol. Chem.* **295**, 4134–4170 (2020).
367. Wang, G. & Vasquez, K. M. Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA Repair* **19**, 143–151 (2014).
368. Bacolla, A. *et al.* Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc. Natl. Acad. Sci.* **101**, 14162–14167 (2004).
369. Bacolla, A., Wojciechowska, M., Kosmider, B., Larson, J. E. & Wells, R. D. The involvement of non-B DNA structures in gross chromosomal rearrangements. *DNA Repair* **5**, 1161–1170 (2006).
370. Carneiro, J., Bernt, M. & Pereira, F. Unravelling the role of non-B DNA conformations in mitochondrial genome rearrangements. in (2017).

371. Carneiro, J., Bernt, M. & Pereira, F. Mitochondrial DNA evolutionary breakpoints are associated with non-B DNA conformations. in (Unpublished, 2018). doi:10.13140/RG.2.2.31991.14242
372. Inagaki, H. *et al.* Chromosomal instability mediated by non-B DNA: Cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans. *Genome Res.* **19**, 191–198 (2009).
373. Wojciechowska, M., Bacolla, A., Larson, J. E. & Wells, R. D. The Myotonic Dystrophy Type 1 Triplet Repeat Sequence Induces Gross Deletions and Inversions. *J. Biol. Chem.* **280**, 941–952 (2005).
374. Wojciechowska, M., Napierala, M., Larson, J. E. & Wells, R. D. Non-B DNA Conformations Formed by Long Repeating Tracts of Myotonic Dystrophy Type 1, Myotonic Dystrophy Type 2, and Friedreich's Ataxia Genes, Not the Sequences per se, Promote Mutagenesis in Flanking Regions. *J. Biol. Chem.* **281**, 24531–24543 (2006).
375. Zhao, J., Bacolla, A., Wang, G. & Vasquez, K. M. Non-B DNA structure-induced genetic instability and evolution. *Cell. Mol. Life Sci.* **67**, 43–62 (2010).
376. Damas, J., Carneiro, J., Amorim, A. & Pereira, F. MitoBreak: the mitochondrial DNA breakpoints database. *Nucleic Acids Res.* **42**, D1261–D1268 (2014).
377. Damas, J., Samuels, D. C., Carneiro, J., Amorim, A. & Pereira, F. Mitochondrial DNA Rearrangements in Health and Disease-A Comprehensive Study. *Hum. Mutat.* **35**, 1–14 (2014).
378. Dong, D. W. *et al.* Association of G-quadruplex forming sequences with human mtDNA deletion breakpoints. *BMC Genomics* **15**, 677 (2014).
379. Markham, N. R. & Zuker, M. in *Bioinformatics* (ed. Keith, J. M.) **453**, 3–31 (Humana Press, 2008).
380. Mukherjee, S. & Bhattacharyya, D. Influence of divalent magnesium ion on DNA: molecular dynamics simulation studies. *J. Biomol. Struct. Dyn.* **31**, 896–912 (2013).
381. Chen, C. & Pettitt, B. M. The Binding Process of a Nonspecific Enzyme with DNA. *Biophys. J.* **101**, 1139–1147 (2011).
382. Young, M. A., Ravishanker, G. & Beveridge, D. L. A 5-nanosecond molecular dynamics trajectory for B-DNA: analysis of structure, motions, and solvation. *Biophys. J.* **73**, 2313–2336 (1997).
383. Pérez, A. *et al.* Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys. J.* **92**, 3817–3829 (2007).
384. Baker, E. S., Dupuis, N. F. & Bowers, M. T. DNA Hairpin, Pseudoknot, and Cruciform Stability in a Solvent-Free Environment. *J. Phys. Chem. B* **113**, 1722–1727 (2009).
385. Auffinger, P. & Westhof, E. Simulations of the molecular dynamics of nucleic acids. *Curr. Opin. Struct. Biol.* **8**, 227–236 (1998).
386. Sagui, C. & Darden, T. A. MOLECULAR DYNAMICS SIMULATIONS OF BIOMOLECULES: Long-Range Electrostatic Effects. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 155–179 (1999).

387. Choi, J. & Majima, T. Conformational changes of non-B DNA. *Chem. Soc. Rev.* **40**, 5893–5909 (2011).
388. Jayaram, B. & Beveridge, D. L. Modeling DNA in Aqueous Solutions: Theoretical and Computer Simulation Studies on the Ion Atmosphere of DNA. *Annu. Rev. Biophys. Biomol. Struct.* **25**, 367–394 (1996).
389. Kar, R. K., Suryadevara, P., Jana, J., Bhunia, A. & Chatterjee, S. Novel G-quadruplex stabilizing agents: in-silico approach and dynamics. *J. Biomol. Struct. Dyn.* **31**, 1497–1518 (2013).
390. Dixit, S. B. *et al.* Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonucleotides. II: Sequence Context Effects on the Dynamical Structures of the 10 Unique Dinucleotide Steps. *Biophys. J.* **89**, 3721–3740 (2005).
391. Brice, A. R. & Dominy, B. N. Analyzing the robustness of the MM/PBSA free energy calculation method: Application to DNA conformational transitions. *J. Comput. Chem.* **32**, 1431–1440 (2011).
392. Carneiro, J., Resende, A. & Pereira, F. The HIV oligonucleotide database (HIVoligoDB). *Database* **2017**, (2017).
393. Carneiro, J., Gomes, C., Couto, C. & Pereira, F. *CoV2ID: Detection and Therapeutics Oligo Database for SARS-CoV-2*. (Bioinformatics, 2020). doi:10.1101/2020.04.19.048991
394. Carneiro, J. & Pereira, F. EbolaID: An Online Database of Informative Genomic Regions for Ebola Identification and Treatment. *PLoS Negl. Trop. Dis.* **10**, e0004757 (2016).
395. *Molecular Modeling of Nucleic Acids*. **682**, (American Chemical Society, 1997).
396. Kramer, M. F. & Coen, D. M. Enzymatic Amplification of DNA by PCR: Standard Procedures and Optimization. *Curr. Protoc. Cytom.* **37**, (2006).
397. Lorenz, T. C. Polymerase Chain Reaction: Basic Protocol Plus Troubleshooting and Optimization Strategies. *J. Vis. Exp.* 3998 (2012). doi:10.3791/3998
398. Owczarzy, R., Moreira, B. G., You, Y., Behlke, M. A. & Walder, J. A. Predicting Stability of DNA Duplexes in Solutions Containing Magnesium and Monovalent Cations. *Biochemistry* **47**, 5336–5353 (2008).
399. Rychlik, W., Spencer, W. J. & Rhoads, R. E. Optimization of the annealing temperature for DNA amplification *in vitro*; *Nucleic Acids Res.* **18**, 6409–6412 (1990).
400. Antczak, M. *et al.* New functionality of RNAComposer: application to shape the axis of miR160 precursor structure. *Acta Biochim. Pol.* **63**, 737–744 (2016).
401. Popena, M. *et al.* Automated 3D structure composition for large RNAs. *Nucleic Acids Res.* **40**, e112–e112 (2012).
402. Jeddi, I. & Saiz, L. Three-dimensional modeling of single stranded DNA hairpins for aptamer-based biosensors. *Sci. Rep.* **7**, 1178 (2017).

403. Heiat, M., Najafi, A., Ranjbar, R., Latifi, A. M. & Rasaee, M. J. Computational approach to analyze isolated ssDNA aptamers against angiotensin II. *J. Biotechnol.* **230**, 34–39 (2016).
404. Sabri, M. Z., Hamid, A. A. A., Hitam, S. M. S. & Rahim, Mohd. Z. A. The assessment of three dimensional modelling design for single strand DNA aptamers for computational chemistry application. *Biophys. Chem.* **267**, 106492 (2020).
405. Schmit, J. D., Kariyawasam, N. L., Needham, V. & Smith, P. E. SLTCAP: A Simple Method for Calculating the Number of Ions Needed for MD Simulation. *J. Chem. Theory Comput.* **14**, 1823–1827 (2018).
406. Ivani, I. *et al.* Parmbsc1: a refined force field for DNA simulations. *Nat. Methods* **13**, 55–58 (2016).
407. Zwanzig, R. Nonlinear generalized Langevin equations. *J. Stat. Phys.* **9**, 215–220 (1973).
408. Portella, G. & Orozco, M. Multiple Routes to Characterize the Folding of a Small DNA Hairpin. *Angew. Chem. Int. Ed.* **49**, 7673–7676 (2010).
409. Goddard, N. L., Bonnet, G., Krichevsky, O. & Libchaber, A. Sequence Dependent Rigidity of Single Stranded DNA. *Phys. Rev. Lett.* **85**, 2400–2403 (2000).
410. Ma, H. T., On, K. F., Tsang, Y. H. & Poon, R. Y. C. An inducible system for expression and validation of the specificity of short hairpin RNA in mammalian cells. *Nucleic Acids Res.* **35**, e22–e22 (2007).
411. McInerney, T. W. *et al.* A globally diverse reference alignment and panel for imputation of mitochondrial DNA variants. *BMC Bioinformatics* **22**, 417 (2021).
412. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160–1166 (2019).
413. Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
414. Brandon, M. C. MITOMAP: a human mitochondrial genome database--2004 update. *Nucleic Acids Res.* **33**, D611–D613 (2004).
415. JASP Team. JASP. (2021). at <<https://jasp-stats.org/>>
416. Plumridge, A., Meisburger, S. P., Andresen, K. & Pollack, L. The impact of base stacking on the conformations and electrostatics of single-stranded DNA. *Nucleic Acids Res.* **45**, 3932–3943 (2017).
417. Yasukawa, T. & Kang, D. An overview of mammalian mitochondrial DNA replication mechanisms. *J. Biochem. (Tokyo)* **164**, 183–193 (2018).
418. Cunha, I., Biltres, R., Sales, M. & Vasconcelos, V. Aptamer-Based Biosensors to Detect Aquatic Phycotoxins and Cyanotoxins. *Sensors* **18**, 2367 (2018).
419. Pereira, A. C. *et al.* Identification of novel aptamers targeting cathepsin B-overexpressing prostate cancer cells. *Mol. Syst. Des. Eng.* **7**, 637–650 (2022).

420. Jiang, Z. *et al.* De Novo Mutation of m.3243A>G together with m.16093T>C Associated with Atypical Clinical Features in a Pedigree with MIDD Syndrome. *J. Diabetes Res.* **2019**, 1–8 (2019).
421. Vellers, H. L. *et al.* Association between Mitochondrial DNA Sequence Variants and V̇O₂ max Trainability. *Med. Sci. Sports Exerc.* **52**, 2303–2309 (2020).
422. Wang, Q. *et al.* Mitochondrial DNA control region sequence variation in migraine headache and cyclic vomiting syndrome. *Am. J. Med. Genet.* **131A**, 50–58 (2004).
423. Azevedo, L. *et al.* Epistatic interactions modulate the evolution of mammalian mitochondrial respiratory complex components. *BMC Genomics* **10**, 266 (2009).
424. Serrano, C. *et al.* Compensatory epistasis explored by molecular dynamics simulations. *Hum. Genet.* **140**, 1329–1342 (2021).
425. Ruan, H., Hao, S., Young, P. & Zhang, H. Targeting Cathepsin B for Cancer Therapies. *Horiz. Cancer Res.* **56**, 23–40 (2015).
426. Gondi, C. S. & Rao, J. S. Cathepsin B as a cancer target. *Expert Opin. Ther. Targets* **17**, 281–291 (2013).
427. Victor, B. C., Anbalagan, A., Mohamed, M. M., Sloane, B. F. & Cavallo-Medved, D. Inhibition of cathepsin B activity attenuates extracellular matrix degradation and inflammatory breast cancer invasion. *Breast Cancer Res.* **13**, R115 (2011).
428. Miyake, H., Hara, I. & Eto, H. Serum level of cathepsin B and its density in men with prostate cancer as novel markers of disease progression. *Anticancer Res.* **24**, 2573–2577 (2004).
429. Abdulla, M.-H. *et al.* Cathepsin B expression in colorectal cancer in a Middle East population: Potential value as a tumor biomarker for late disease stages. *Oncol. Rep.* **37**, 3175–3180 (2017).
430. Kotaska, K., Dusek, P., Prusa, R., Vesely, S. & Babjuk, M. Urine and Serum Cathepsin B Concentrations in the Transitional Cell Carcinoma of the Bladder: Cathepsin B in Bladder Cell Carcinoma. *J. Clin. Lab. Anal.* **26**, 61–65 (2012).
431. Heidtmann, H.-H. *et al.* Cathepsin B and cysteine proteinase inhibitors in human lung cancer cell lines. *Clin. Exp. Metastasis* **15**, 368–381 (1997).
432. Ruan, J. *et al.* Over-expression of cathepsin B in hepatocellular carcinomas predicts poor prognosis of HCC patients. *Mol. Cancer* **15**, 17 (2016).
433. Sun, T. *et al.* Expression profile of cathepsins indicates the potential of cathepsins B and D as prognostic factors in breast cancer patients. *Oncol. Lett.* **11**, 575–583 (2016).
434. Yang, W.-E. *et al.* Cathepsin B Expression and the Correlation with Clinical Aspects of Oral Squamous Cell Carcinoma. *PLOS ONE* **11**, e0152165 (2016).
435. Ferreira, D. *et al.* Selection of aptamers against triple negative breast cancer cells using high throughput sequencing. *Sci. Rep.* **11**, 8614 (2021).
436. Zhou, J. & Rossi, J. Aptamers as targeted therapeutics: current potential and challenges. *Nat. Rev. Drug Discov.* **16**, 181–202 (2017).

437. Khot, V. M. *et al.* Nanomedicine-driven molecular targeting, drug delivery, and therapeutic approaches to cancer chemoresistance. *Drug Discov. Today* **26**, 724–739 (2021).
438. Fattal, E., Hillaireau, H. & Ismail, S. I. Aptamers in Therapeutics and Drug Delivery. *Adv. Drug Deliv. Rev.* **134**, 1–2 (2018).
439. Sousa, D., Ferreira, D., Rodrigues, J. L. & Rodrigues, L. R. in *Appl. Target. Nano Drugs Deliv. Syst.* 357–409 (Elsevier, 2019). doi:10.1016/B978-0-12-814029-1.00014-4
440. Pereira, A., Sales, M. & Rodrigues, L. in *Adv. Biosens. Health Care Appl.* 71–103 (Elsevier, 2019). doi:10.1016/B978-0-12-815743-5.00003-2
441. Bayat, P. *et al.* SELEX methods on the road to protein targeting with nucleic acid aptamers. *Biochimie* **154**, 132–155 (2018).
442. Cerchia, L. Aptamers: Promising Tools for Cancer Diagnosis and Therapy. *Cancers* **10**, 132 (2018).
443. Morita, Y., Leslie, M., Kameyama, H., Volk, D. & Tanaka, T. Aptamer Therapeutics in Cancer: Current and Future. *Cancers* **10**, 80 (2018).
444. Fu, Z. & Xiang, J. Aptamers, the Nucleic Acid Antibodies, in Cancer Therapy. *Int. J. Mol. Sci.* **21**, 2793 (2020).
445. Li, Z. *et al.* Advances in Screening and Development of Therapeutic Aptamers Against Cancer Cells. *Front. Cell Dev. Biol.* **9**, 662791 (2021).
446. Bavi, R., Liu, Z., Han, Z., Zhang, H. & Gu, Y. In silico designed RNA aptamer against epithelial cell adhesion molecule for cancer cell imaging. *Biochem. Biophys. Res. Commun.* **509**, 937–942 (2019).
447. Wang, Y. *et al.* Label-free microfluidic paper-based electrochemical aptasensor for ultrasensitive and simultaneous multiplexed detection of cancer biomarkers. *Biosens. Bioelectron.* **136**, 84–90 (2019).
448. Bell, D. R. *et al.* In silico design and validation of high-affinity RNA aptamers targeting epithelial cellular adhesion molecule dimers. *Proc. Natl. Acad. Sci.* **117**, 8486–8493 (2020).
449. Wang, Q.-L., Cui, H.-F., Du, J.-F., Lv, Q.-Y. & Song, X. In silico post-SELEX screening and experimental characterizations for acquisition of high affinity DNA aptamers against carcinoembryonic antigen. *RSC Adv.* **9**, 6328–6334 (2019).
450. Santini, B. L. *et al.* In Silico Design of Novel Mutant Anti-MUC1 Aptamers for Targeted Cancer Therapy. *J. Chem. Inf. Model.* **60**, 786–793 (2020).
451. Rockey, W. M. *et al.* Rational Truncation of an RNA Aptamer to Prostate-Specific Membrane Antigen Using Computational Structural Modeling. *Nucleic Acid Ther.* **21**, 299–314 (2011).
452. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
453. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**, 129–137 (1982).

454. MacQueen, J. B. Some methods for classification and analysis of MultiVariate observations. in *Proc. 5th Berkeley Symp. Math. Stat. Probab.* (eds. Cam, L. M. L. & Neyman, J.) **1**, 281–297 (University of California Press, 1967).
455. Roel-Touris, J., Don, C. G., V. Honorato, R., Rodrigues, J. P. G. L. M. & Bonvin, A. M. J. J. Less Is More: Coarse-Grained Integrative Modeling of Large Biomolecular Assemblies with HADDOCK. *J. Chem. Theory Comput.* **15**, 6358–6367 (2019).
456. Rodrigues, J. P. G. L. M. *et al.* Clustering biomolecular complexes by residue contacts similarity. *Proteins Struct. Funct. Bioinforma.* **80**, 1810–1817 (2012).
457. Miller, B. R. *et al.* MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **8**, 3314–3321 (2012).
458. Wang, E. *et al.* End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chem. Rev.* **119**, 9478–9508 (2019).
459. Sgobba, M., Caporuscio, F., Anighoro, A., Portioli, C. & Rastelli, G. Application of a post-docking procedure based on MM-PBSA and MM-GBSA on single and multiple protein conformations. *Eur. J. Med. Chem.* **58**, 431–440 (2012).
460. Shen, M. *et al.* Discovery and optimization of triazine derivatives as ROCK1 inhibitors: molecular docking, molecular dynamics simulations and free energy calculations. *Mol. Biosyst.* **9**, 361–374 (2013).
461. Zhang, X., Perez-Sanchez, H. & C. Lightstone, F. A Comprehensive Docking and MM/GBSA Rescoring Study of Ligand Recognition upon Binding Antithrombin. *Curr. Top. Med. Chem.* **17**, 1631–1639 (2017).
462. Santos-Pereira, C. *et al.* The milk-derived lactoferrin inhibits V-ATPase activity by targeting its V1 domain. *Int. J. Biol. Macromol.* **186**, 54–70 (2021).
463. Onufriev, A., Bashford, D. & Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins Struct. Funct. Bioinforma.* **55**, 383–394 (2004).
464. Jokar, M., Safaralizadeh, M. H., Hadizadeh, F., Rahmani, F. & Kalani, M. R. Design and evaluation of an apta-nano-sensor to detect Acetamidiprid *in vitro* and *in silico*. *J. Biomol. Struct. Dyn.* **34**, 2505–2517 (2016).
465. Aamir, M. *et al.* In silico Prediction, Characterization, Molecular Docking, and Dynamic Studies on Fungal SDRs as Novel Targets for Searching Potential Fungicides Against Fusarium Wilt in Tomato. *Front. Pharmacol.* **9**, 1038 (2018).
466. Suenaga, A., Okimoto, N., Hirano, Y. & Fukui, K. An Efficient Computational Method for Calculating Ligand Binding Affinities. *PLoS ONE* **7**, e42846 (2012).
467. Kelly, L., Maier, K. E., Yan, A. & Levy, M. A comparative analysis of cell surface targeting aptamers. *Nat. Commun.* **12**, 6275 (2021).
468. Gopal, P., Rehman, R. U., Chadha, K. S., Qiu, M. & Colella, R. Matrigel influences morphology and cathepsin B distribution of prostate cancer PC3 cells. *Oncol. Rep.* **16**, 313–320 (2006).
469. Podgorski, I. *et al.* Bone Microenvironment Modulates Expression and Activity of Cathepsin B in Prostate Cancer. *Neoplasia* **7**, 207–223 (2005).

470. Poreba, M. *et al.* Fluorescent probes towards selective cathepsin B detection and visualization in cancer cells and patient samples. *Chem. Sci.* **10**, 8461–8477 (2019).
471. Graycar, T. P., Bott, R. R., Power, S. D. & Estell, D. A. in *Handb. Proteolytic Enzym.* 3148–3155 (Elsevier, 2013). doi:10.1016/B978-0-12-382219-2.00693-1
472. Lemkul, J. A. & MacKerell, A. D. Polarizable Force Field for DNA Based on the Classical Drude Oscillator: I. Refinement Using Quantum Mechanical Base Stacking and Conformational Energetics. *J. Chem. Theory Comput.* **13**, 2053–2071 (2017).
473. Kurkcuoglu, Z. & Bonvin, A. M. J. J. Pre- and post-docking sampling of conformational changes using ClustENM and HADDOCK for protein-protein and protein-DNA systems. *Proteins Struct. Funct. Bioinforma.* **88**, 292–306 (2020).
474. Neves, R. P. P., Fernandes, P. A. & Ramos, M. J. Role of Enzyme and Active Site Conformational Dynamics in the Catalysis by α -Amylase Explored with QM/MM Molecular Dynamics. *J. Chem. Inf. Model.* **62**, 3638–3650 (2022).

F. APPENDIX

This page was intentionally left blank.

1. Posters

This page was intentionally left blank.

1.1. Application of molecular dynamics simulations in the study of the PLP synthase macrocomplex

BImBS school 2019 – BioInformatics meets BioSimulations in protein and DNA studies: from theory to practice, 2019, Lugano, Switzerland



Biomolecular SIMulations Research Group
Info@biosim.pt www.biosim.pt

U.PORTO
FACULDADE DE MEDICINA
UNIVERSIDADE DO PORTO

cost
EUROPEAN COOPERATION
SCIENTIFIC & TECHNICAL

cecam

Application of molecular dynamics simulations in the study of the PLP synthase macrocomplex

André F. Pina*, Sérgio F. Sousa, and Nuno M.F.S.A. Cerqueira

UCIBIO@REQUIMTE, BioSIM, Department of Biomedicine - Faculty of Medicine, University of Porto, Porto, Portugal

* apina@med.up.pt in/andrefpina @andrefpina

INTRODUCTION

Tuberculosis and malaria are responsible for the annual death of millions of people worldwide. A major factor hindering the control of these diseases is the high degree of resistance developed by these pathogens against the currently available drugs.^[1]

The *de novo* DXP-independent biosynthesis of vitamin B₁₂ is usually seen as a possible therapeutic target since, while it is absent in humans, its inhibition impairs the development of *Mycobacterium tuberculosis* and *Plasmodium falciparum*.^[2,4]

This pathway is composed by a remarkable macromolecular complex, the **PLP synthase**, a cogwheel-like protein formed by 12 synthase subunits to which 12 glutaminase subunits attach, the Pdx1 and Pdx2, respectively. Alone, it can synthesize the cofactor pyridoxal 5'-phosphate (PLP) directly from glyceraldehyde 3-phosphate, ribose 5-phosphate, and glutamine.^[3]

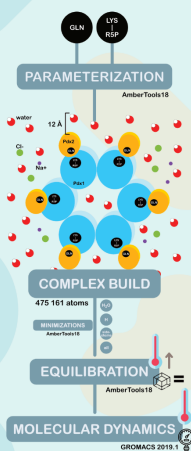
In this work, we aim to study the conformational dynamics underlying the function of PLP synthase, through the use of molecular dynamics simulations.

[1] World Health Organization. 'Global tuberculosis report 2018'. World Health Organization, Geneva (2018).
[2] World Health Organization. 'World malaria report 2018'. World Health Organization, Geneva (2018).
[3] Dik, U., Marjono, B., Kijne, M., Gengenbacher, M., Microbiol., 78, 180 (2010).
[4] G. Guadalupe, V. Hays, V. Kishore, S. Ganes, M. Gengenbacher, B. Kishore, J. Schilling, J. Tamm, S. Kishore, S. E. Enk, T. P. Bagley, J. Tamm, Nat. Chem. Biol., 13, 200 (2017).

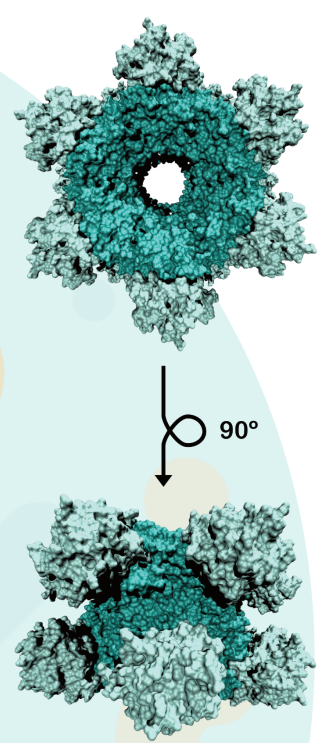
METHODOLOGY

The crystallographic structure of plasmodial PLP synthase, deposited on Protein Data Bank with the code 4ADS³⁴ was used to prepare the model for the molecular dynamics simulation. The PLP synthase system consists of 12 Pdx1 monomers, 12 Pdx2 monomers, 12 L-glutamine molecules, 12 ribose 5-phosphate molecules covalently bonded to the protein, water, and 150 mM NaCl.

The procedure used is described in the image below:



The system minimizations and equilibration phase were performed in AmberTools18³⁵ with the force field ff15FB in combination with TIP3P-FB water molecules.³⁷ The topology files were then converted to GROMACS format using ParmEd. The molecular dynamics simulation was performed using GROMACS 2019.1.³⁶ An integration time step of 2 fs was used with Verlet algorithm.³⁸ The PME algorithm was used for electrostatic interactions.³⁹ The Verlet cutoff-scheme with 1.0 nm was used for short range electrostatic and van der Waals interactions. The V-rescale⁴⁰ and the Parrinello-Rahman⁴¹ algorithms were used for temperature and pressure coupling.

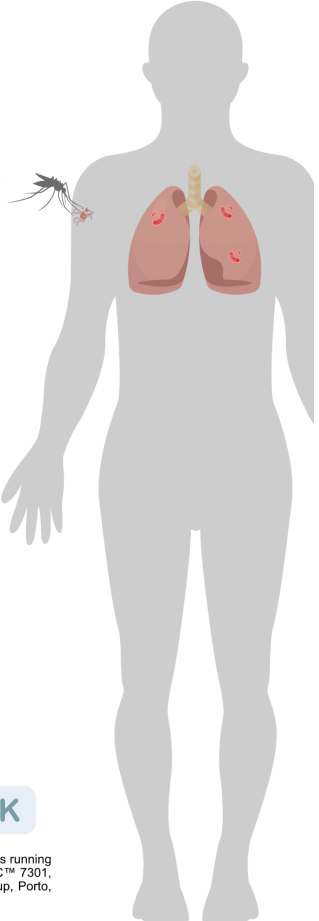


ONGOING WORK

Currently, the molecular dynamics simulation is running on a cluster of dual CPU sockets, AMD EPYC™ 7301, in the FLASH HPC at BioSIM Research Group, Porto, Portugal.

FUTURE PERSPECTIVES

This knowledge will help us to understand the tight influence that the quaternary structure of PLP synthase has in its catalytic process, guiding us in the construction of representative QM/MM models for the study of its catalytic mechanism.



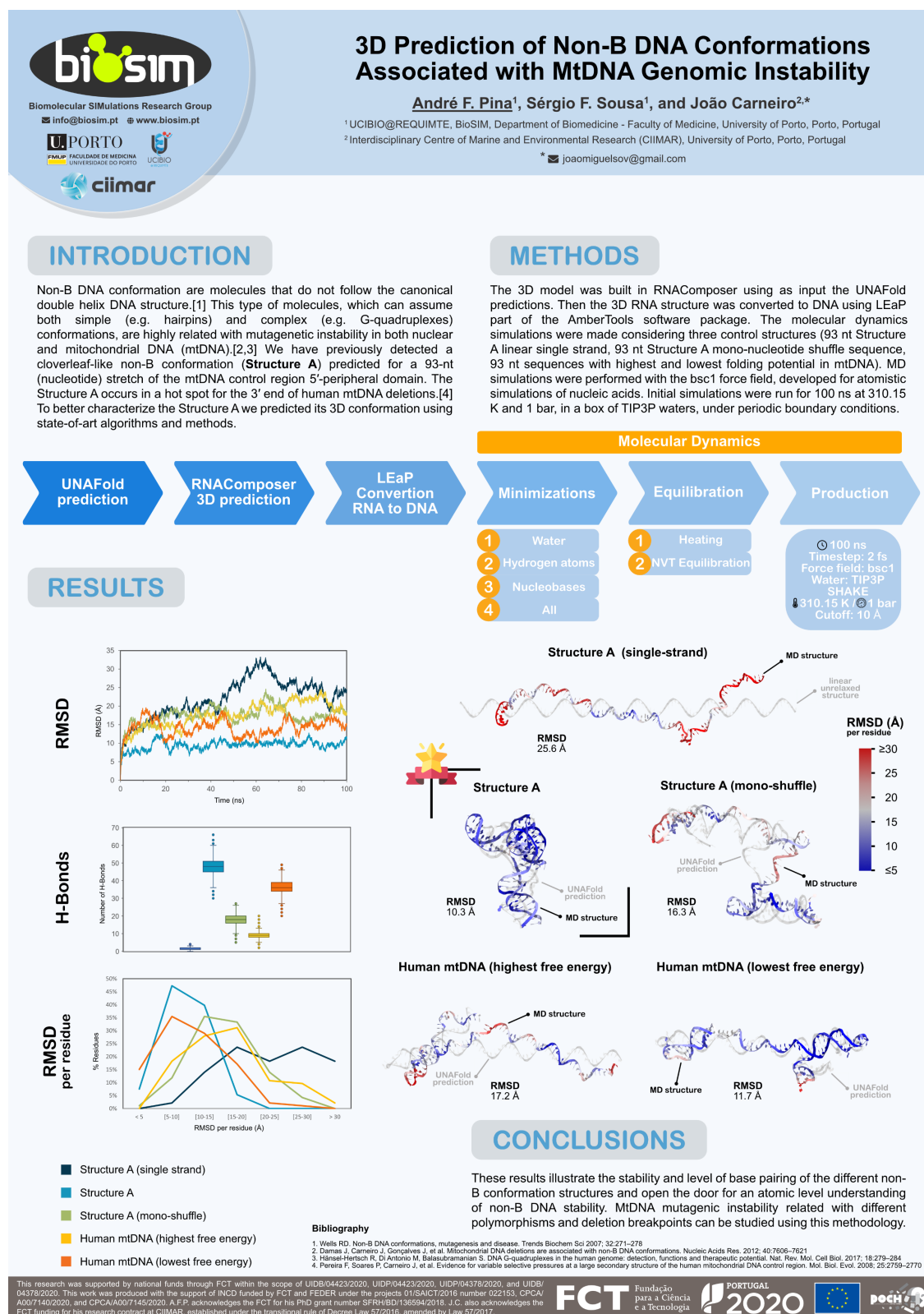
This work was supported by national funds from Fundação para a Ciência e a Tecnologia (SFRH/BD/136594/2018, IF/00052/2014, PTDC/QUI-QF/31680/2017, and UID/Multi/04378/2019) and co-financed by the ERDF under the PT2020 Partnership Agreement (POCI-01-0145-FEDER-007728).



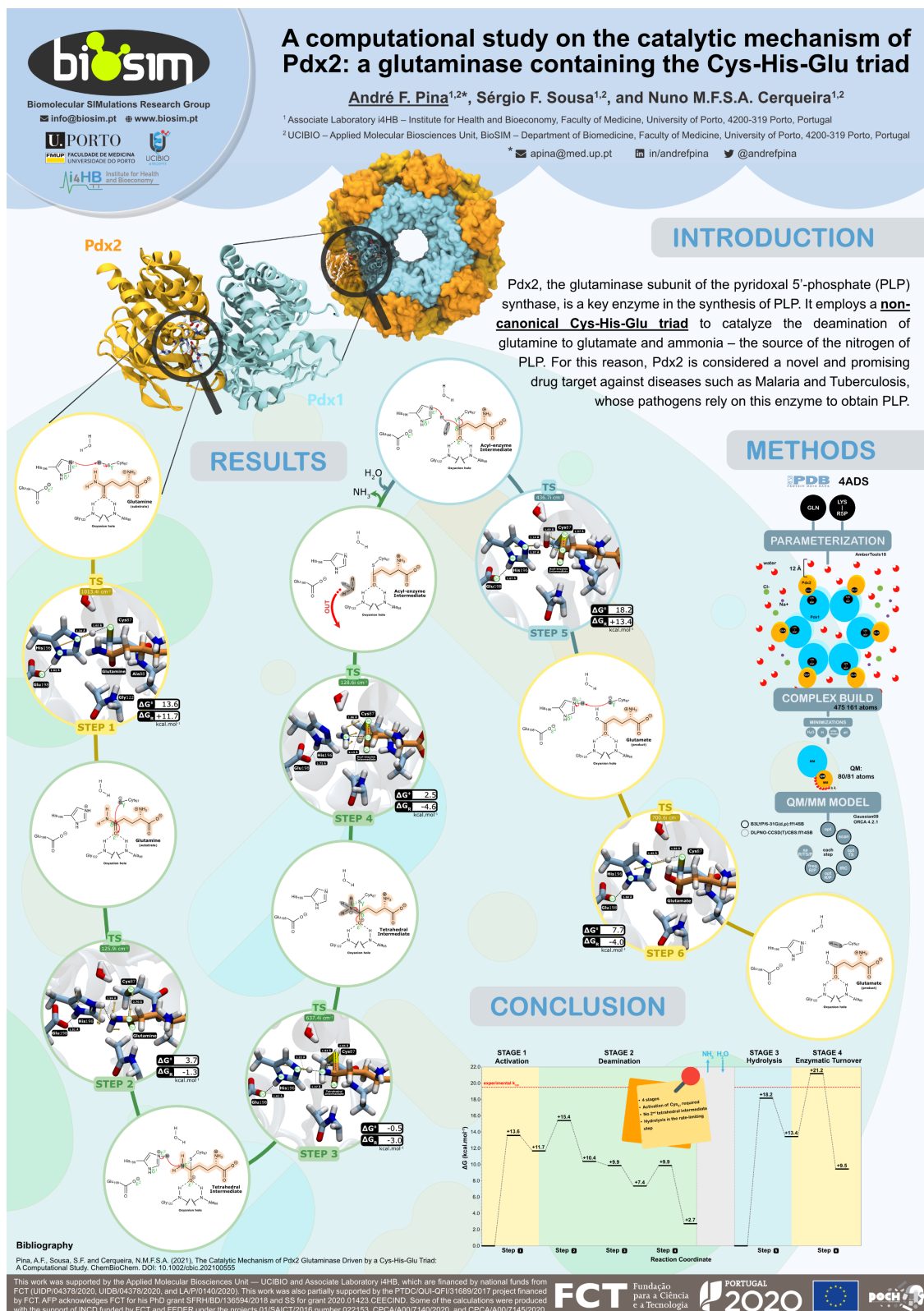


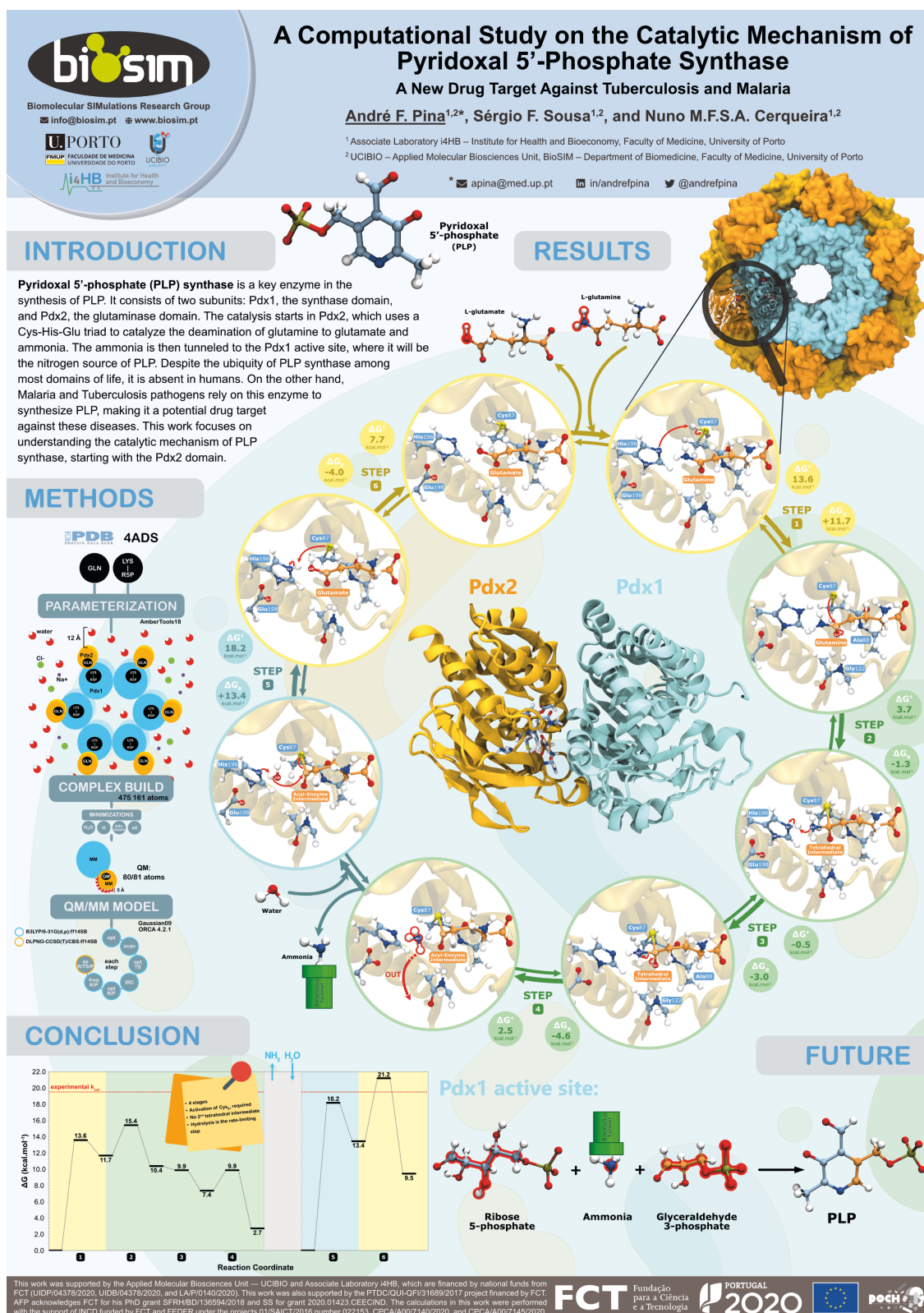

1.2. 3D prediction of non-B DNA conformations associated with mtDNA genomic instability

Bioinformatics Open Days, 2021, online



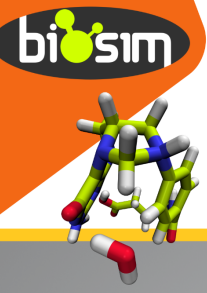
VIII EJIBCE – Encontro De Jovens Investigadores Em Biologia Computacional Estrutural,
2021, Coimbra, Portugal





1.5. Development of computational methods for aptamer design for biomedical applications

5th Meeting in Medicinal Biotechnology / 2nd Iberian Congress in Medicinal Biotechnology, 2022, Porto, Portugal



Development of Computational Methods for Aptamer Design for Biomedical Applications

André F. Pina, Sérgio Filipe Sousa

UCIBIO@REQUIMTE, BioSIM, Departamento de Biomedicina, Faculdade de Medicina, Universidade do Porto, Portugal

sergiofsousa@med.up.pt | www.biosim.pt






molUP-A VMD plugin to handle QM and ONIOM calculations using the gaussian software
Fernandes HF et al. (2018), 39, 1344

Challenges

THE DEFINITION
Aptamers are short single-stranded nucleic acids or peptide molecules that can bind with high affinity to specific protein targets.

They have been employed as a very specific tool for molecular recognition or to targeted drug delivery.

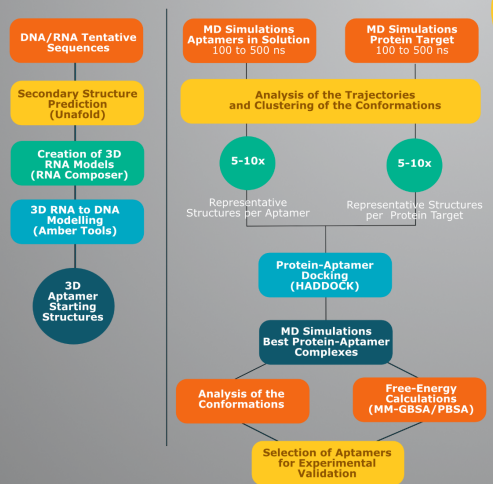
Over the last decades hundreds of aptamers have been selected to recognize different molecular targets, including small molecules, peptides and specific proteins. Many have been patented for specific applications.

EXPERIMENTAL SELECTION
Aptamer selection is normally accomplished by a process of "in vitro selection" named SELEX, which means systematic evolution of ligands by exponential enrichment.

In this process randomized libraries of different sequences are subject to iterative cycles of selection/separation/amplification, yielding round after round sequences with a higher affinity towards a specific target.

IN SILICO METHODS FOR RATIONAL DESIGN OF APTAMERS
More recently, in silico methods have been developed to help understand aptamer-target interaction and to rationally introduce modifications in selected aptamers to modulate their affinity, specificity or ability to carry other molecules. In addition, in silico efforts to design aptamers with high affinity to specific targets, reducing the time and cost for developing rapid diagnostics tools and effective therapeutics has been gaining momentum.

The Approach



Objectives

- Rational Design and Selection of Aptamers
- Understand Protein-Aptamer Interaction
- Improve Specificity
- Module Properties

Recent Publications

Biocatalysis

The Catalytic Mechanism of Threonine Aldolase, a Critical Enzyme in the Pharmaceutical Industry to Synthesize β-Hydroxy-α-amino Acids
Rocha J et al. ACS Catalysis 12 (9) pp. 4904-4909 (2022)

The critical role of Asp108 in the catalytic mechanism of the threonine aldolase PpTase
Machado R et al. Catalysis Science and Technology 10 (20) pp. 10350-10357 (2020)

The Catalytic Mechanism of PpG Glucanase Driven by a Cys-His-Asp Catalytic Triad: A Computational Study
Pina A et al. ChemRxiv 2021.09.10.458100 (2021)

An Unusual Cys-Glu-His Catalytic Triad is Responsible for the Catalytic Mechanism of the Nucleoside Sphingomyelinase: A QM/MM Study on MD
Rocha J et al. ChemRxiv 2021.09.10.458100 (2021)

Catalytic Mechanism of the Serine Hydroxymethyltransferase: A Computational QM/MM Study
Fernandes H et al. ACS Catalysis 11 (11) pp. 10366-10370 (2021)

Mechanistic Studies of a From Monocysteine Sulfur Cluster of Oligonucleotides by DsC
Barbosa A et al. ACS Catalysis 11 (10) pp. 5038-5051 (2021)

Unravelling the Catalytic Mechanism and the Reaction Pathway of the Malonaldehyde Transaminase Domain of Human Fatty Acid Synthase
Pina A et al. ACS Catalysis 11 (9) pp. 4860-4872 (2021)

Improving the Catalytic Power of the DsC Enzyme for the Biocatalysis of Cyclic Olefins and Derivatives
Ferreira J et al. Chemistry - A European Journal 26 (38) pp. 10721-10731 (2020)

Case Study

Identification of novel aptamers targeting cathepsin B-overexpressing prostate cancer cells

Mol. Syst. Des. Eng., 2022, 7, 637

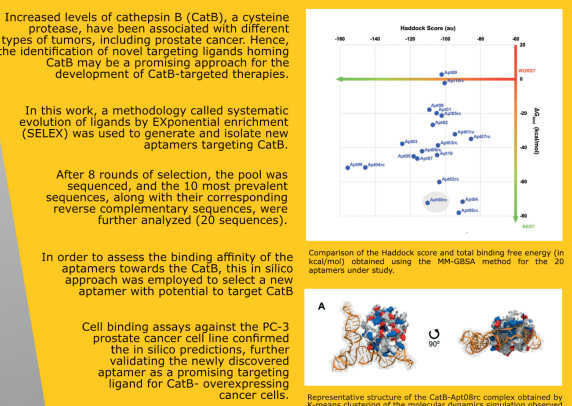
Increased levels of cathepsin B (CatB), a cysteine protease, have been associated with different types of tumors, including prostate cancer. Hence, the identification of novel targeting ligands homing CatB may be a promising approach for the development of CatB-targeted therapies.

In this work, a methodology called systematic evolution of ligands by EXponential enrichment (SELEX) was used to generate and isolate new aptamers targeting CatB.

After 8 rounds of selection, the pool was sequenced, and the 10 most prevalent sequences, along with their corresponding reverse complementary sequences, were further analyzed (20 sequences).

In order to assess the binding affinity of the aptamers towards the CatB, this in silico approach was employed to select a new aptamer with potential to target CatB

Cell binding assays against the PC-3 prostate cancer cell line confirmed the in silico predictions, further validating the newly discovered aptamer as a promising targeting ligand for CatB-overexpressing cancer cells.



Other Covers



VMD Store

A VMD Plugin to Browse, Discover, and Install VMD Extensions

Fernandes H.S. et al. Journal of Chemical Information and Modeling 59 (11) pp.4519-4523 (2019)

Assessing the effects of PM60 variants on protein stability
Guthrie D. et al. Molecular Genetics and Metabolism 134 (4) pp.344-352 (2021)

The role of the malonaldehyde transaminase (MALT) domain in targeting of H1 histone
Santos-Pereira C. et al. International Journal of Biological Macromolecules 180 pp.54-70 (2021)

Conformational aptamers designed by molecular dynamics simulations
Santos C. et al. Human Genetics 140 (9) pp.1359-1362 (2021)



www.biosim.pt

This work was supported by national funds from the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding UIDB/04378/2020 and UIDP/04378/2020. The author acknowledges FCT by funding 2020.01423 CEECIND/CP1596/CT0003.



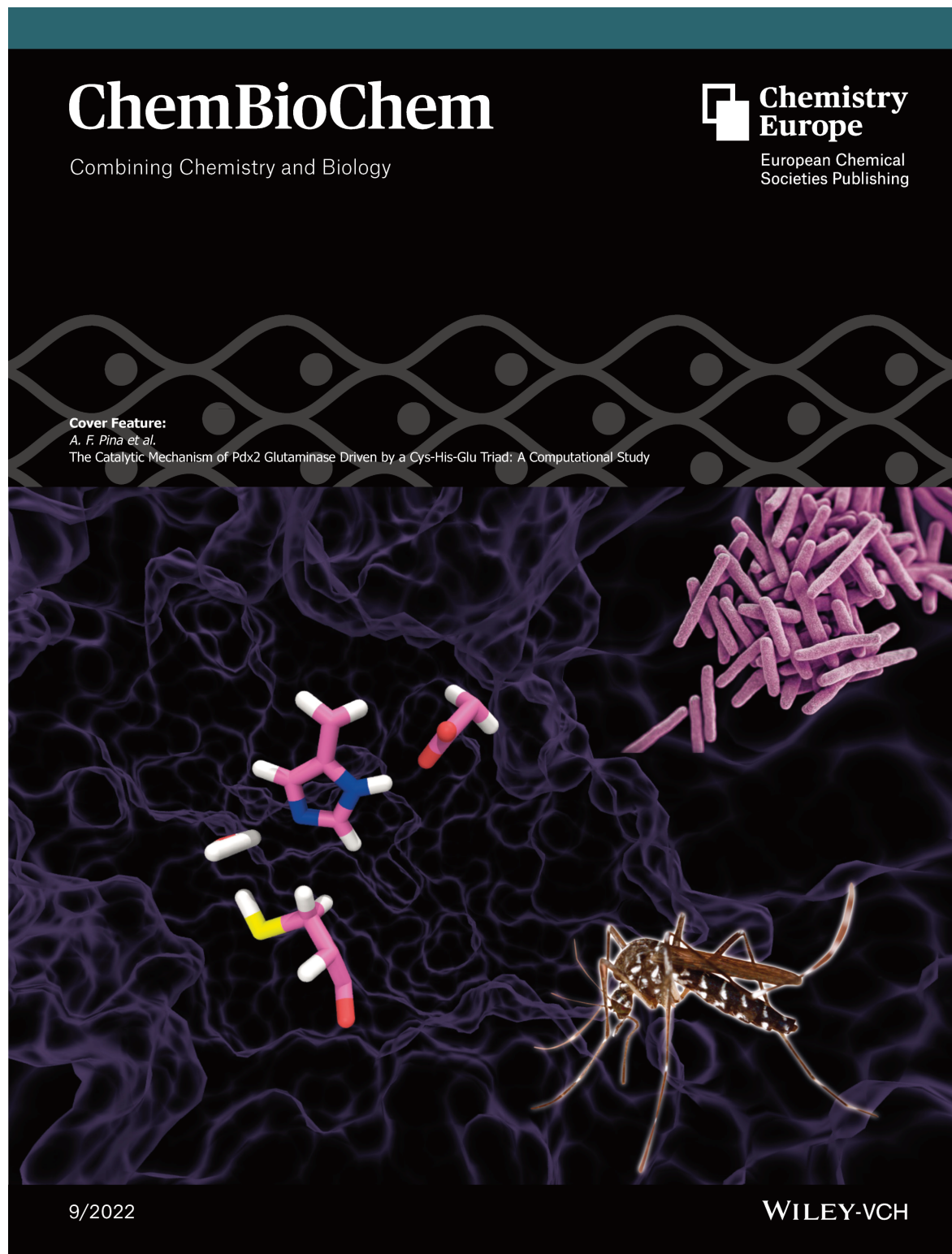

This page was intentionally left blank.

2. Covers

This page was intentionally left blank.

2.1. Cover of the ChemBioChem

Volume 23, Issue 9, Page e202200024 (2022)



This page was intentionally left blank.

3. Reviews

This page was intentionally left blank.



Cite this: *Catal. Sci. Technol.*, 2019, 9, 4864

PLP-dependent enzymes as important biocatalysts for the pharmaceutical, chemical and food industries: a structural and mechanistic perspective

Juliana F. Rocha,  André F. Pina, 
Sérgio F. Sousa  and Nuno M. F. S. A. Cerqueira *

Enzymes catalyze a plethora of chemical reactions in nature. These biocatalysts offer a variety of benefits that no other synthetic catalyst can offer: high turnover, regio- and stereoselectivity, and ability to work under mild conditions and to react in environmentally friendly manufacturing processes. However, enzymes are not normally optimally suited for industrial applications. It is important to improve their properties to enable their more efficient application in an industrial setting, while also expanding the range of potential chemical reactions that they can catalyze. In this review, we focus our analysis on structural features and mechanisms of different enzymes that contain a pyridoxal-5'-phosphate (PLP) cofactor and that have noteworthy potential as biocatalysts in a variety of industries and applications. The selected PLP enzymes are ω -transaminases, lysine decarboxylase, threonine aldolase, L-tyrosine phenol-lyase, α -amino- ϵ -caprolactam racemases, and cystathionine β -lyase. For each of these enzymes, the catalytic mechanism, industrial application, and the advantages and disadvantages of their application are reviewed in detail. In general, this review highlights the immense biocatalytic potential, rich chemistry and diverse set of applications that different enzymes sharing a common element (PLP) can offer.

Received 19th June 2019,
Accepted 9th August 2019

DOI: 10.1039/c9cy01210a

rsc.li/catalysis

Introduction

Enzymes are perhaps the primary biocatalysts used by industrial biotechnology (IB) since they can increase the rate of a reaction by several orders of magnitude. Their production is readily available from renewable sources. In addition, enzymes are also biodegradable and essentially nontoxic and non-hazardous. Furthermore, the reactions catalyzed by enzymes are normally performed under mild conditions (ambient temperature and atmospheric pressure) in an aqueous environment, contrasting markedly with the typical chemical industrial processes, which require high operating temperatures and pressures, or potentially toxic organic solvents, and are expensive. In addition, the use of enzymes often reduces the number of steps required in the reactions, avoiding protection and deprotection steps. The application of enzymes as biocatalysts represents, therefore, a more cost-effective, environmentally friendly and sustainable alternative, with potential application in a wide range of chemical diverse industrial problems.^{2,3}

Despite the potential offered by enzymes as biocatalysts, there is still hesitation on their use since they are not optimally suited for industrial applications. Additionally, their products compete with products from the petroleum-based industry. This industry is one of the most important sources for chemical products and energy. However, the finite nature of oil reserves, coupled with increasing concerns about the environmental damage caused by man, urges us to develop more sustainable ways of securing these chemicals and energy.⁴ It is therefore important to expand the range of chemical reactions that enzymes can catalyze and improve their properties. The optimization of the activity and substrate specificity of enzymes is still expensive and difficult to achieve by experimental means, as this involves understanding at the atomic level the catalytic mechanism of these enzymes^{5–8} and rationally designing new ones^{9,10} with the desired properties and activities.

PLP-dependent enzymes have an unequalled catalytic versatility, making them extraordinary biocatalysts. They use pyridoxal-5'-phosphate (PLP) as a cofactor to catalyze a diverse range of reactions that are specifically controlled by the active site of the enzyme that contains it.¹¹ PLP resembles benzaldehyde in its structure. However, the properties of the carbonyl group are modified by the presence both of an

UCIBIO/REQUIMTE, BioSIM, Departamento de Biomedicina, Faculdade de Medicina, Universidade do Porto, Alameda Professor Hernâni Monteiro, 4200-319 Porto, Portugal. E-mail: nunoscerqueira@med.up.pt

adjacent hydroxyl group and ring nitrogen in the *para* position. This allows PLP to very easily form imines with primary amino groups through its aldehyde group, and to stabilize carbanionic intermediates formed by the heterolytic cleavage of chemical bonds. A wide range of reactions has already been found, such as racemization, decarboxylation, beta addition, beta-elimination, retro-aldol cleavage, transamination, and alpha-elimination, among others.^{12,13} Despite the diversity of reactions that are catalyzed by PLP-dependent enzymes, it is possible to envisage a series of common mechanistic features characteristic of PLP-catalyzed reactions.

Structurally, PLP-dependent enzymes can be classified into 7 different fold-types: fold-types I to V as suggested by Grishin *et al.*¹⁴ (Fig. 1), and two additional fold-types VI and VII that contain some enzyme families that could not be categorized into the original fold-types.¹⁵ A recent detailed update regarding the diversity of fold-types of PLP-dependent enzymes has been presented by Liang *et al.*¹⁶ The enzymes reviewed in this article belong to fold types I to IV.

These enzymes are only active when the PLP cofactor is covalently bonded to a lysine residue in the active site, generating what is commonly known as internal aldimine.^{17,18} When the substrate is available, the transamination reaction occurs, in which PLP becomes covalently bonded to the substrate, forming what is commonly known as external aldimine.¹⁹ From this point, the reaction diverges, and it is different for different types of PLP-dependent enzymes. Once the product of the reaction is formed, it remains covalently bonded to the PLP cofactor as the external aldimine. The product of the reaction is only released from the active site to the solvent, after the transamination reaction (the reverse of the first step), from which the PLP cofactor becomes again covalently bonded to the active site lysine residue and forms an internal aldimine. From this point, the PLP enzymes are ready for a new turnover.

In this context, the PLP-dependent enzymes selected as examples in this review – ω-transaminases, lysine decarboxylase, threonine aldolase, L-tyrosine phenol-lyase, α-amino-ε-caprolactam racemase, and cystathionine β-lyase – have be-

come interesting targets for different industrial fields (Scheme 1), due to the chemical versatility of PLP-dependent enzymes in the catalysis of a wide variety of chemical reactions, with high regio- and stereoselectivity. This makes them powerful tools in the semi-synthetic production of natural amino acids, non-canonical amino acids and their related compounds, which are required in many industrial processes. How different members of the PLP enzyme family are able to accept a variety of substrates is vitally important to understand the use of these enzymes in industrial applications, a fundamental feature to improve their activity.

Transaminases (TAmS)

Chiral amines are valuable and versatile building blocks for the pharmaceutical industry since it is estimated that 40% of all pharmaceuticals contain an amine functionality.²⁰ Due to increasing environmental awareness, the development of new biocatalysts for the production of optically active amines is of the utmost interest. One group of enzymes that has been showing great potential for the asymmetric synthesis of chiral amines is transaminases (TAmS).^{21–23} As biocatalysts, they offer a green

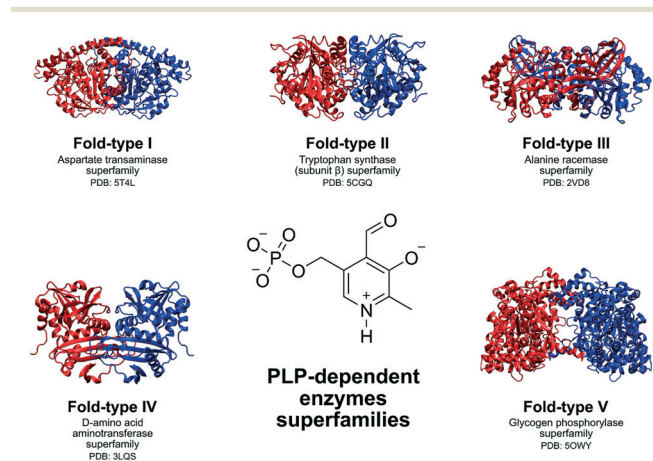
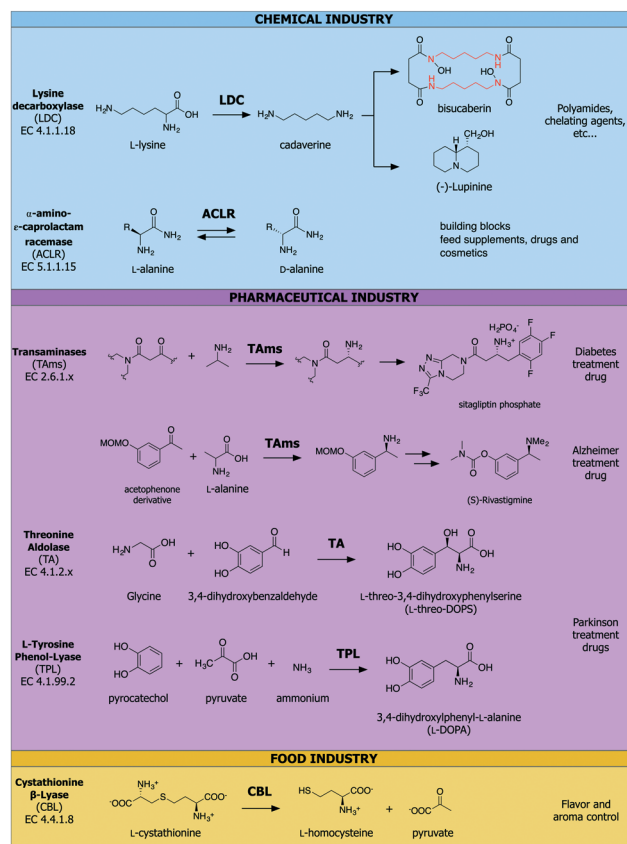


Fig. 1 3D structures representative of the main fold-types of each PLP-dependent superfamily.



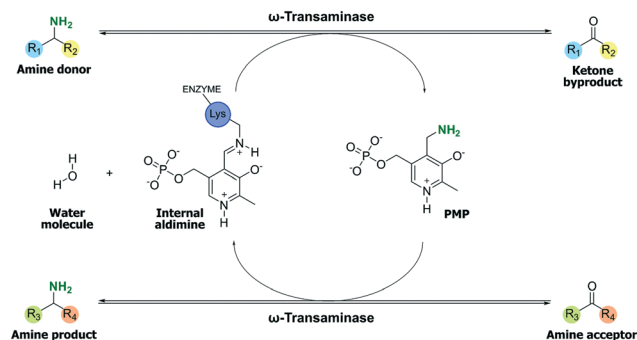
Scheme 1 Schematic representation of industrial applications of the different PLP-dependent enzymes presented in this review. In blue, LDC and ACLAR are presented which have application in the chemical industry; in purple, TAmS, TA and TPL are presented which have application in the pharmaceutical industry; and in yellow, CBL is presented which has application in the food industry.

and economically viable alternative to transition metal catalysts in the stereoselective amination of prochiral ketones.²⁴

Transaminases (EC 2.6.1.X), also known as aminotransferases, are PLP-dependent enzymes responsible for the transference of an amino group from an amino donor to an acceptor (Fig. 2). PLP, the cofactor, acts as the transient keeper of the amino group, being bound as a Schiff-base to a lysine residue in the active site (the internal aldimine).¹⁸

The transamination reaction can be considered as two different half-reactions (Scheme 2). In the first half-reaction, the amine group of an amine donor (either an amino acid or another amine) replaces the lysine residue in the internal aldimine, originating from the external aldimine. Then, the first half of the reaction culminates in the generation of pyridoxamine 5'-phosphate (PMP) and the corresponding ketone. In the second half-reaction, the amine group now present in PMP is transferred to an amine acceptor (ketone, keto acid or aldehyde) with the regeneration of the internal aldimine and formation of the corresponding amine, completing the catalytic cycle.

According to the type of substrate that is converted, transaminases (TAMs) can be divided into two groups: α -transaminases (α -TAMs) and ω -transaminases (ω -TAMs). α -TAMs require the presence of a carboxylic group in the α -position with respect to the carbonyl functionality.²⁰ They are involved in the biosynthesis of α -amino acids, performing the transamination of the corresponding α -keto acids.²⁵ On the other hand, ω -TAMs are able to accept aliphatic ketones and amines as their substrates (*i.e.* not only α -keto acids and amino acids), representing all the remaining TAMs.²⁴ ω -TAMs have been further subdivided into two subgroups, β -TAMs²⁶ and amine transaminases (ATAMs), the latter often used erroneously as a synonym for all ω -TAMs. This terminol-



Scheme 2 Representation of the overall reaction catalyzed by ω -transaminases.

ogy can be misleading as although ATAMs can convert any amine substrate independently of the presence of a carboxylic group, not all ω -TAMs are able to do so.²⁷

As PLP-dependent enzymes, TAMs can also be classified based on their fold-type, belonging to fold-type I (aspartate transaminase superfamily) and fold-type IV (α -amino acid aminotransferase superfamily).²⁸ In addition to their classification based on their fold type, transaminases are additionally divided into six classes according to their similarity.²⁷

One of the main characteristics of ω -TAMs is their high enantiospecificity and relatively broad substrate range—useful properties in an industrial context. As previously mentioned, ω -TAMs are important tools in the synthesis of valuable building blocks for pharmaceutical production, such as chiral amines. It has been shown that ω -TAMs can be used as catalysts in kinetic resolution and asymmetric synthesis of optically pure amines.²⁴

In nature, (*S*)-selective TAMs are more abundant than the (*R*)-selective ones, so, for a very long time, mainly (*S*)-selective ω -TAMs were known and extensively studied.²⁹ Thus, in order to prepare (*R*)-amines, the only green method available was the kinetic resolution of racemic amines by (*S*)-selective TAMs, a method that only allows a maximum yield of 50 percent.³⁰ Hence, there was an urgent need for the discovery of (*R*)-selective TAMs to access a broad range of (*R*)-amines by the efficient asymmetric synthesis, a method that can reach a theoretical yield of 100 percent and has become the preferred method for the pharmaceutical industry.

In the last decade, several (*R*)-selective ω -TAMs have been described and are being studied in regard to their substrate scope.^{29,31–38} Several engineering approaches have been applied in the expansion of the substrate scope of (*R*)-selective ω -TAMs towards bulky substrates.^{39–43} The synthetic potential of (*R*)-selective TAMs has been highlighted by the success of its application in the large-scale production of the anti-diabetic drug sitagliptin in 2010. The use of an engineered (*R*)-selective TAM from *Arthrobacter* sp. (ATA 117) came to replace the previously applied asymmetric chemical hydrogenation, a rhodium-catalyzed process, in the conversion of a bulky ketone.^{44–46} This new biocatalytic route led to a drastic increase in yield (10–13%) and productivity (53%, measured

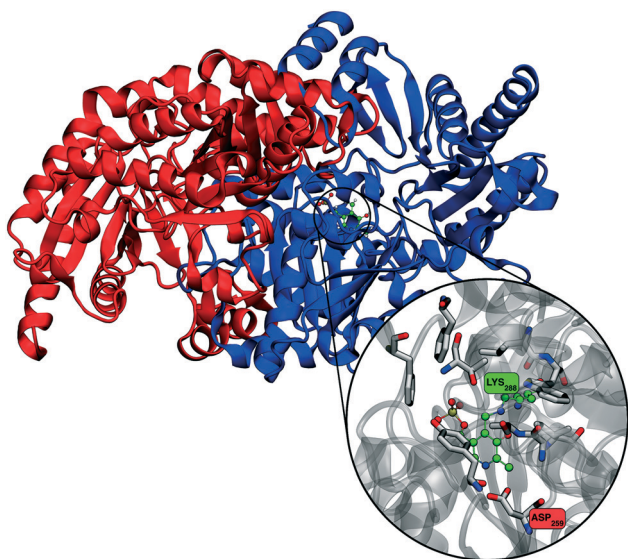


Fig. 2 3D structure of ω -TAM from *Chromobacterium violaceum* (PDB code: 4A6T). Some of the important residues in the active site are highlighted in the circle, together with the internal aldimine which is represented in green.

in kg L^{-1} per day), being also responsible for the decrease of total waste (19%), and the need for heavy metals, while reducing the total manufacturing costs.⁴¹ This was a milestone that has inspired others to put their hands to new engineering projects.^{40,42,47}

The determination of more and more structures of (*S*)-selective TAm and (*R*)-selective TAm has enabled a deeper investigation about the underlying mechanism of these enzymes, as well as their substrate specificity and/or stereospecificity.

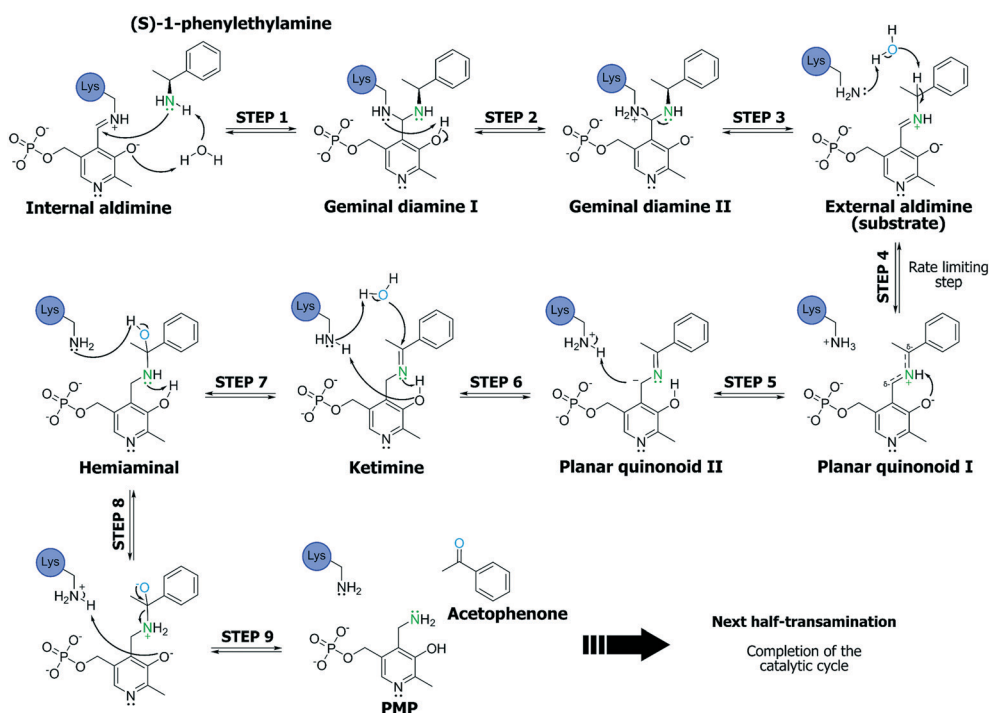
Their active site consists of a large-binding pocket and a small-binding pocket, that together form the substrate-binding site, achieved by dimerization of the two monomers that compose the enzyme.⁴⁰ Commonly, the small binding pocket is able to accommodate only a methyl group, thus revealing the extreme relevance of the engineered TAm to accept bulkier substituents.³²

A unique ability that characterizes TAm is the capacity to accept both hydrophilic and hydrophobic substrates in the same active site. It was found that, in (*S*)-selective TAm, the cause for this is due to the presence of a “flipping” arginine residue in a loop near the active site. Its guanidinium group is able to move in the active site and can interact either with the carboxylate group of the substrate or move towards the outside, providing a hydrophobic space for the bulkier substrate in that active site region.⁴⁸ Moreover, Guan *et al.* found that in (*R*)-selective TAm, the arginine residue is present in only one conformation, contrary to what happens in (*S*)-selective TAm. Nevertheless, the whole loop containing the arginine residue is able to move, thus acting as a substrate-specificity determinant, making it a key area when

attempting to modify the substrate specificity through protein engineering.⁴⁹

The detailed mechanism of ω -TAm has not been completely established. However, it has been assumed in the literature that their mechanism must be similar to that of aspartate transaminase.^{50–52} A common first step in all PLP-dependent enzymes is the formation of a Schiff-base of PLP and an amino substrate, called the external aldimine, which replaces the lysine residue in the internal aldimine.

One study of a ω -TAm reaction mechanism (Scheme 3) was performed computationally by Cassimjee *et al.* through the application of density functional theory calculations.¹ The crystallographic structure used (PDB 4A6T) is an (*S*)-selective ω -TAm purified from the bacteria *Chromobacterium violaceum*. The work focused on the half-transamination of (*S*)-1-phenylethylamine to acetophenone. The results suggest that the amino substrate forms an external aldimine with the cofactor pyridoxal 5'-phosphate (PLP) through geminal diamine intermediates. The rate-limiting step of the reaction is the deprotonation of the external aldimine, with an energy barrier of $19.3 \text{ kcal mol}^{-1}$, culminating in the formation of a planar quinonoid intermediate. Then, a ketimine is formed, followed by the production of a hemiaminal through the addition of a water molecule. Finally, a ketone product is obtained together with the formation of pyridoxamine 5'-phosphate (PMP). Upon dissociation of the ketone from the active site, it becomes available for the binding of the second substrate, thus becoming ready for the next half-reaction, with the reformation of the internal aldimine (enzyme-PLP) and the completion of the catalytic cycle. These results



Scheme 3 Representation of the mechanism proposed by Cassimjee *et al.* [adapted].¹

indicate that ω -TAMs might follow the same overall mechanism as other transaminases.

There is an ever-growing demand for the development of new biocatalysts for the synthesis of enantiopure amines, an important role that can be played by ω -TAMs. Successful cases such as the application of an engineered TAM in the industrial synthesis of sitagliptin only validate the power and efficiency of this biocatalyst. However, the full potential of TAMs is still hampered by a restricted substrate scope, co-product inhibition, and difficulties in displacing reaction equilibria. For these reasons, ω -TAMs still do not have the impact as a synthetic approach for the pharmaceutical industry. Nevertheless, the accelerated rate of discovery of new ω -TAMs, especially the (*R*)-selective ones, and the increasing protein engineering approaches towards these enzymes are a clear sign that ω -TAMs will continue to develop as a competitive alternative technology to the conventional synthesis methods currently used.

Lysine decarboxylase (LDC)

Cadaverine is a molecule of great interest for industrial applications since it has a similar structure to the synthesized petrochemical polymers, most notably hexamethylenediamine. It can, therefore, be used as an alternative to such compounds in the production of polyamides or polyurethanes.^{53,54} Cadaverine has already been used successfully as a precursor for the production of alkaloids, polyamides, polyurethanes, chelating agents, and additives.^{55–57} As classical petrochemical polymers have a global market of approximately 7 million tons per year, their replacement by bio-polymers would be an important step forward in the implementation of IB.⁵⁶

One way to produce cadaverine is by using the enzyme lysine decarboxylase (LDC). LDC catalyzes the decarboxylation of L-lysine to produce cadaverine with the capture of a proton and the release of carbon dioxide.^{56,58} Lysine decarboxylase (EC 4.1.1.18) is a fold-type I PLP-dependent enzyme (Fig. 3) that is widely distributed in prokaryotes and eukaryotes.⁵⁹ In the literature, it is possible to identify two types of lysine decarboxylase: a constitutive and an inducible one.^{53,58} They have very similar amino acid sequences despite having difference in the optimum pH and thermal stability.⁵³ The most used for cadaverine production is the inducible form since it exhibits higher thermal and catalytic stability.⁵³

The efficient use of LDC on an industrial scale has been hampered mainly due to the inhibition of this enzyme by its end-product,⁵⁹ and its limited efficiency outside a narrow pH range.⁶⁰ When the concentration of cadaverine is at 3 g L⁻¹, the catalytic activity of the enzyme drops to half.⁵⁹ The dependence of the catalytic activity on the pH is related to structural conformations. While a decamer is the favorable structural conformation at an optimum pH of 6, when the reaction proceeds, the pH of the medium increases since hydrogen ions are being consumed. At a lower pH, the enzyme usually forms oligomers and, at higher pH values, decamers and dimers are formed, with the percentage of dimers in-

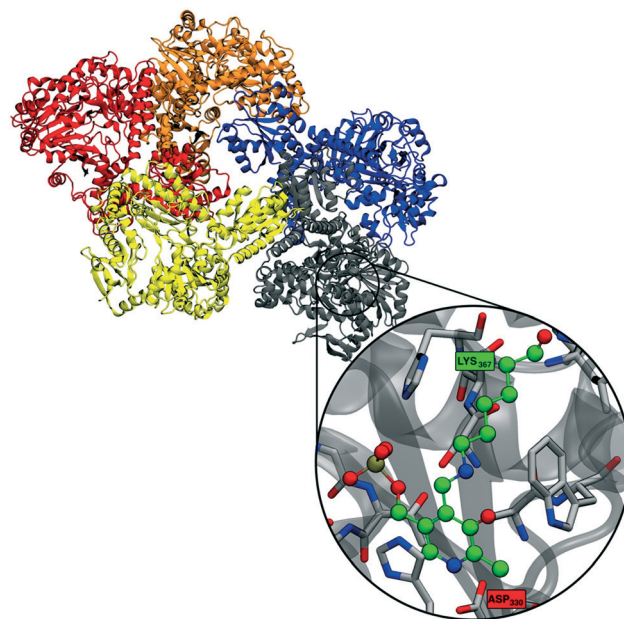


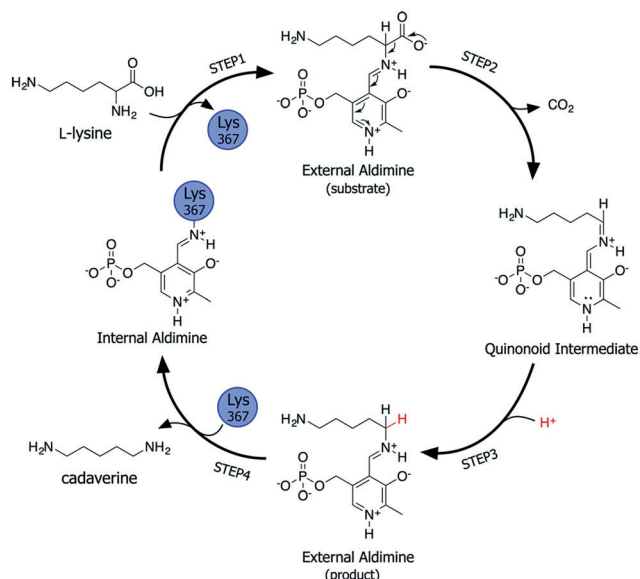
Fig. 3 3D structure of LDC (PDB code: 3N75). In the bottom, the active site of the enzyme is presented. The internal aldimine is represented in green.

creasing with pH. Dimers have a lower specific activity than decamers.

Hence, the control of the structure of the enzyme, and of the pH of the medium can be used as methods to improve the biosynthesis of cadaverine. One solution to this problem is the development of an engineered enzyme. In order to stabilize the structure of the enzyme, Park *et al.* tried to immobilize it.⁵⁶ This study showed some improvements since the free enzyme cannot be recovered after being used, while the immobilized one was recovered and, after 10 cycles, still showed 53 percent residual activity.⁵⁶ The immobilization rate was, however, very low, so, further studies are needed.

Although the mechanism of this enzyme is not described in the literature, it is thought to be very similar to that of other decarboxylases. A recent study on histidine decarboxylase¹² suggests that the mechanism occurs in two sequential steps: in the first step, the decarboxylation of L-lysine occurs, followed by a second step where the protonation takes place. By extrapolating, we can suggest a mechanism for lysine decarboxylase (Scheme 4), in which this whole process is only possible due to the presence of PLP. Initially, PLP is linked to a lysine residue (internal aldimine). In the presence of L-lysine, PLP binds to it forming an external aldimine (step 1). Decarboxylation of L-lysine is now promoted and a quinonoid intermediate is formed, with the release of carbon dioxide (step 2). Upon protonation of the quinonoid intermediate with a proton taken from the medium (step 3), the internal aldimine is re-established with the release of the final product, cadaverine (step 4).

Cadaverine biosynthesis through LDC is a promising route. However, to be regarded as an efficient alternative,



Scheme 4 The suggested mechanism for lysine decarboxylase.

further studies are needed to overcome the limitation caused by the fact that the product inhibits the enzyme.

Threonine aldolase (TA)

β -Hydroxy- α -amino acids (HAAs) are core building blocks of several drugs, such as antibiotics, immunosuppressants, and the anti-Parkinson's disease drug L-threo-3,4-dihydroxyphenylserine (L-threo-DOPS).^{61–63} However, the industrial production of HAA is expensive and partly inefficient from a chemical point of view due to the two chiral centers formed upon synthesis, which results in a mixture of stereoisomers.⁶⁴ The enzyme threonine aldolase (TA) catalyzes the retro-aldol cleavage of threonine into glycine and acetaldehyde, and also the reversible reaction where the condensation of acetaldehyde and glycine occurs to form threonine with two chiral centers.⁶¹ Due to its reversibility, TA is a powerful catalyst for asymmetric carbon–carbon bond formation, enabling the production of a wide range of HAAs from glycine and aldehyde.

Threonine aldolases are PLP-dependent enzymes (Fig. 4) which are widely distributed in nature.⁶⁵ They can be classified according to their specificity at α -carbon and β -carbon. These enzymes are highly specific at α -carbon and they can be divided into L-TA or D-TA. L-TA can be divided into three categories according to the specificity of the β -carbon: highly specific L-threonine aldolases (EC 4.1.2.5), which only accept L-threonine; L-*allo*-threonine aldolases (EC 4.1.2.49), which accepts the rare stereoisomer L-*allo*-threonine; and L-low-specificity-threonine aldolase (EC 4.1.2.48) that can accept both L- and L-*allo*-threonine.⁶⁵ Until now, only D-low-specificity-threonine aldolase (EC 4.1.2.42) has been found in nature and can accept both D- and D-*allo*-threonine.

L-TA and D-TA have distinct structures and different folding types. L-TA (fold-type I) functions as a homotetramer and the active site is located at the dimer interface between sub-

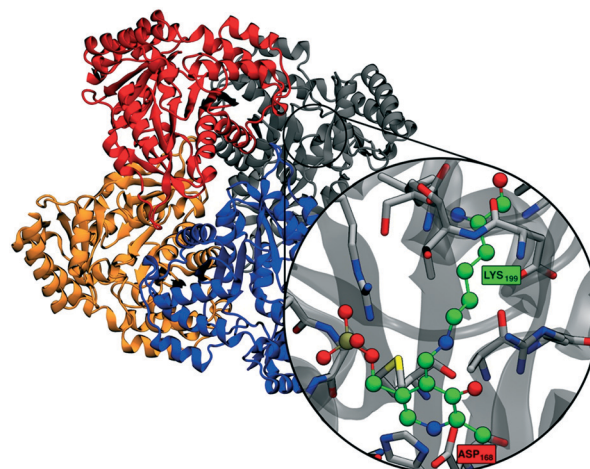


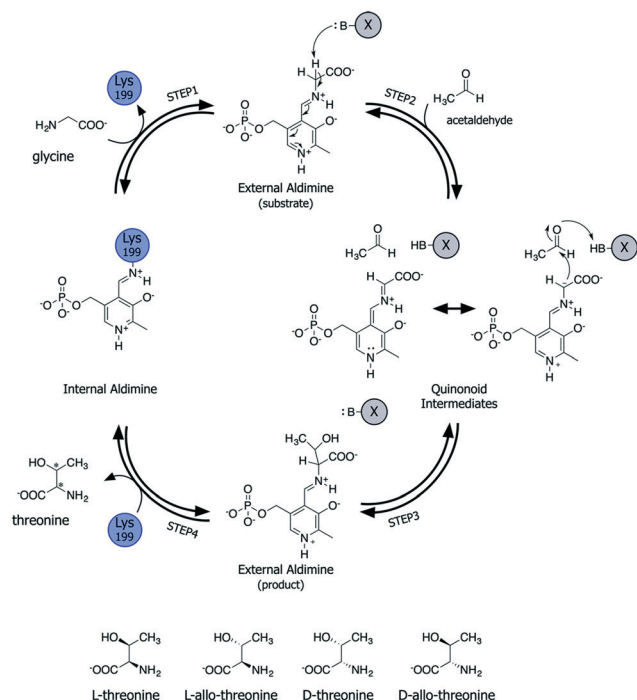
Fig. 4 3D structure of L-TA (PDB code: 1M6S). In the bottom, the active site of the enzyme is presented. The internal aldimine is represented in green.

units,⁶⁶ while D-TA (fold-type III) was isolated as a dimer.⁶⁷ Notwithstanding, crucial catalytic residues seem to have a similar orientation on both TA.

The low-specificity of TA on β -carbon has hampered their use in industrial processes. Enzyme engineering has been attempted to overcome this limitation, including mutagenesis studies^{62,68–72} and immobilization techniques.^{73,74} These attempts have tried to increase the stereospecificity and thermostability of the enzyme. However, so far, the outcome is still far from the desired goals.

Although the mechanism of TA is still not characterized, it is proposed to occur as follows (Scheme 5): initially, PLP is linked to a catalytic lysine residue (internal aldimine). When glycine is available, PLP releases the lysine residue and becomes covalently bonded to glycine forming an external aldimine (step 1). The external aldimine promotes the deprotonation of the α -carbon of glycine by a residue that acts as a general base (step 2). This step leads to a quinonoid intermediate, with high resonance, that allows the formation of a carbanion ready for a nucleophilic attack. Since acetaldehyde is a good acceptor for the nucleophilic attack, due to its electrophilicity, when it enters the active site, it allows the formation of a C–C bond with the quinonoid intermediate, with the capture of a proton from the general base (step 3). For the enzymatic turnover, PLP is transferred from the product to the catalytic lysine, reestablishing the internal aldimine and releasing the final product, threonine (step 4). This general mechanism is thought to be followed by different TAs, resulting in similar products with variations in stereochemistry (Scheme 5).

TAs are promising enzymes for the biosynthesis of HAAs. However, their reaction rates and stereoselectivity continue to be poorly optimized for industrial use. Even with several improvements of enzymatic engineering already achieved, further studies are needed in order for TA to be considered, from an industrial point of view, as an efficient alternative for the biosynthesis of HAAs to the traditional methods.



Scheme 5 Suggested mechanism for threonine aldolases. Different stereochemistry structures of threonine are presented (* refers to chiral centres in the molecule).

L-Tyrosine phenol-lyase (TPL)

3,4-Dihydroxyphenyl-L-alanine (L-DOPA) is the drug of choice for the treatment of Parkinson's disease. Annually, more than 250 tons of L-DOPA are produced, and half of it comes from enzymatic processes.⁷⁵ The extraction of L-DOPA from natural sources⁷⁶ no longer satisfies the actual demand, and the chemical route employs so extreme reaction conditions that it is not an efficient option either.⁷⁷

L-DOPA can be biosynthesized using L-tyrosine phenol-lyase (TPL) from different organisms.⁷⁵ TPL catalyzes the reversible cleavage of the C β –C γ bond from L-tyrosine to form phenol, pyruvate, and ammonium.⁷⁸ Since this reaction is reversible, TPL can be used for the production of L-DOPA in one step from common chemicals.⁷⁹

TPL from *Erwinia herbicola* has been regarded, for some years now, as the most efficient alternative for the production of L-DOPA. However, large-scale production of L-DOPA through TPL is still not efficient in part due to the fact that the expression of TPL from *Erwinia herbicola* in *E. coli* decreases, significantly, the reaction rates.⁸⁰ Recently, Zheng *et al.* described TPL from *Fusobacterium nucleatum* that has a comparable production rate of L-DOPA,⁸⁰ but is easier to express.

TPL (EC 4.1.99.2) is a bacterial PLP-dependent enzyme (fold-type I) found in a homotetramer arrangement (Fig. 5), with active sites located at each monomer–monomer interface, totaling four active sites.^{81–83} These enzymes have a characteristic that has been revealed to be crucial for the catalytic activity: a rearrangement of the domains allows the ac-

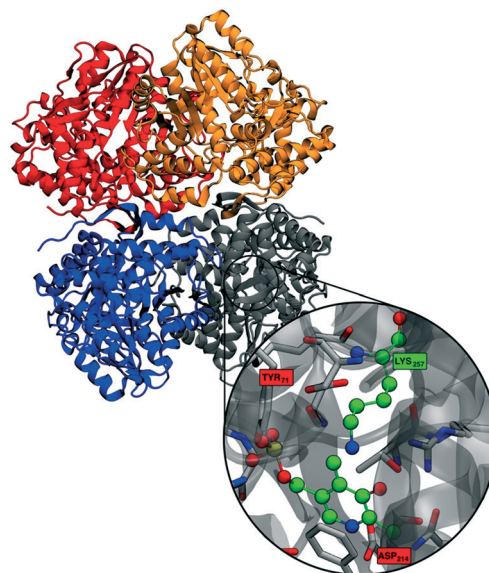
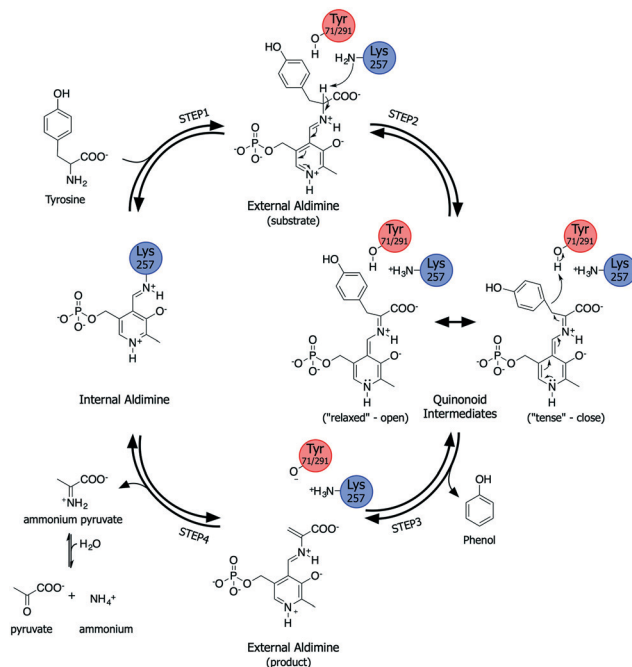


Fig. 5 3D structure of TPL (PDB code: 1C7G). In the bottom, the active site of the enzyme is presented. The internal aldimine is represented in green.

tive site to adopt two different conformations – an open and a closed conformation.⁸²

The proposed mechanism for TPL involves several steps (Scheme 6), in which the formation of a “tense” form of the quinonoid intermediate is crucial for the catalytic cycle.^{82,84} Similar to other PLP-dependent enzymes, the first step involves the formation of an external aldimine in the presence of L-tyrosine, with the cleavage of the internal aldimine. The catalytic lysine acts then as a general base, deprotonating the



Scheme 6 The suggested mechanism for L-tyrosine phenol-lyase.

$C\alpha$ of the substrate, forming a quinonoid intermediate in a relaxed form (step 2). The quinonoid intermediate undergoes a rearrangement, adopting a “tense” conformation which favors the protonation of $C\gamma$ through a tyrosine residue and the cleavage of the $C\beta$ – $C\gamma$ bond releasing the first product, phenol (step 3). This rearrangement is crucial for this step but it is still not clear if the protonation and bond cleavage occur in a stepwise or in a concerted way (here it is represented in a concerted way). At last, the internal aldimine is reestablished and ammonium pyruvate is released (step 4), completing the catalytic cycle. Additionally, the ammonium pyruvate reacts with water leading to the formation of pyruvate and ammonium.

L-DOPA is a first-line treatment drug for Parkinson's disease. Nowadays, this disease affects increasingly more people due to the aging of the population, which increases the demand. Wild-type TPL is still not efficient for large-scale industrial production, but efforts are being made to find efficient solutions to optimize this enzyme, in order to make it a biocatalyst of excellence.

α -Amino- ε -caprolactam racemases (ACLRs)

D-Amino acids and their derivatives are of extreme importance in the chemical industry, being building blocks for other compounds with biological or physiological activities, as well as, products with direct commercial value, being applied in feed supplements, drugs, and cosmetics. However, while proteinogenic amino acids can be easily synthesized by microbial fermentation, other amino acids, such as D-isomers, require other approaches since they are scarcely produced by fermentation.⁸⁵ Therefore, there is a need for new and efficient synthetic pathways for the production of enantiopure amino acids in the chemical industry.

An elegant method that has been increasingly applied for the synthesis of these compounds is the enzyme-mediated dynamic kinetic resolution (DKR) of amino acids and their derivatives. One example is the enantioselective enzymatic hydrolysis of racemic amino acid amides (AAAs) by a D-stereospecific hydrolase, such as D-amino peptidase. This enzyme hydrolyses all D-AAAs into D-amino acids while leaving L-AAAs untouched. This allows a maximum theoretical yield of 50 percent if only the hydrolase is present in the reaction mixture. However, if an AAA racemase is also present in the system, all the remaining L-AAAs will be racemized into D-AAAs, allowing the complete hydrolysis of the racemic mixture into the desired product. Moreover, D-amino acids can be produced by this method. The synthesis of L-amino acids is also possible if the D-stereospecific hydrolase is exchanged with an L-stereospecific one.^{86–88}

An excellent racemase biocatalyst that can be used in the DKR of AAAs is the PLP-dependent enzyme, α -amino- ε -caprolactam (ACL) racemase (ACLR, EC 5.1.1.15, Fig. 6). This enzyme is responsible for the interconversion of L- and D-ACL, and has been used in the industrial synthesis of L-lysine.^{89,90} It

was initially reported by Ahmed *et al.* that the only substrates of ACLR were ACL, α -amino- δ -valerolactam, and α -amino-3-thio- ε -caprolactam.^{89,91} However, Asano and Yamaguchi discovered that ACLR also has racemization activity, not only towards cyclic AAAs, but towards acyclic AAAs.^{86,87} More recently, a study published by Frese *et al.* described the discovery of an ACLR enzyme with promiscuous activity towards amino acid esters (AAEs). Despite its low AAE activity, it provides a robust basis for the engineering of ACLRs towards improved performance in dynamic resolution processes.⁹²

Its crystallographic structure was first determined by Okazaki *et al.* for the *Achromobacter obae* ACLR (AoACLR), revealing that ACLR is a dimeric enzyme, that belongs to the fold-type I group of PLP-dependent enzymes (Fig. 6).⁹³ Its active site is located at the dimeric interface, being constituted by residues from both monomers.

ACLR catalyzes the removal of the α -hydrogen from one face of the substrate and the addition of a proton on the opposite side, therefore achieving the racemization of the substrate. The substrate binding site was typically located between Trp49 and Tyr137, as suggested by the position of the ε -caprolactam moiety in the complex structure obtained by Okazaki *et al.* for the AoACLR.⁹³

Two mechanisms have been proposed for this reaction: a single-base mechanism and a two-base mechanism. The single-base mechanism was proposed by Ahmed *et al.* based on a series of deuterium labelling experiments, suggesting that Lys267 was the presumable base responsible for the protonation of the planar quinonoid intermediate from either side.⁹⁴

However, recent evidence found by Okazaki *et al.* suggests that the racemization occurs through a two-base mechanism, one in each face of the substrate (Scheme 7). The proposed residues involved in this process are Lys267 (from the *re*-face)

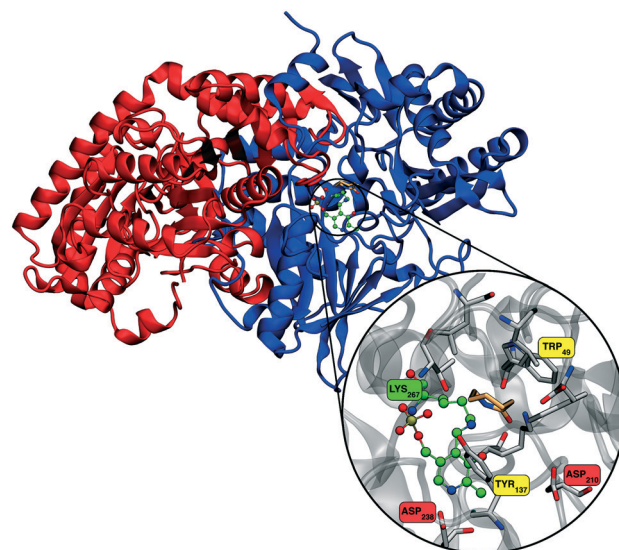
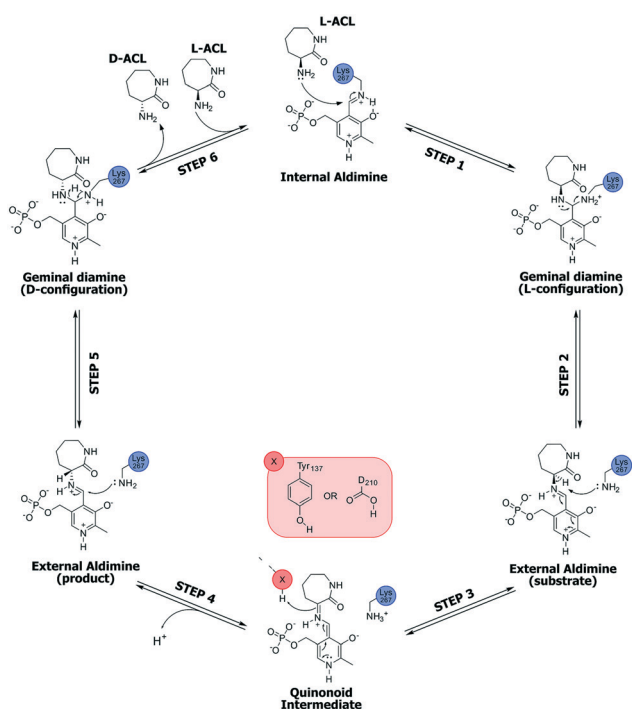


Fig. 6 3D structure of ACLR from *Achromobacter obae* (PDB code: 3DXW). In the bottom, the active site of the enzyme is presented. The internal aldimine is represented in green and the ε -caprolactam moiety in orange.

and Tyr137 (from the *si*-face).⁹³ This is emphasized by the work performed by Awad and his co-workers where they compared several fold-type I PLP-dependent enzymes with racemization activity, including the ACLR from *Achromobacter obae*.⁹⁵ In fact, a mutagenic study was performed using the enzyme cystathionine β -lyase, a fold-type I PLP-dependent enzyme with a structure similar to ACLR. When Tyr111 (Tyr 137 in ACLR) is mutated by phenylalanine, its racemization activity towards L-alanine is lost, backing up the hypothesis that Lys and Tyr are the catalytic pair.⁹⁶

Nevertheless, there has been some controversy about this matter. According to their crystallographic study, Frese *et al.* suggested that the candidate acid/base catalytic residues are Lys267 and Asp210, not Tyr137.⁹⁷ Although, due to technical reasons, they were not able to obtain the Y137 mutant, basing their conclusions only on the poor activity of the D210A mutant, a residue that albeit its importance, may not be part of the catalytic pair.

The growing use of chiral amino acids, as building blocks in industry, dictates that new ways must be found to produce them efficiently and in a single enantiomer form. A key enzyme involved in this process is the ACLR, achieving that through the dynamic kinetic resolution of AAAs. The more complete understanding of its mechanism, together with the discovery of ACLRs with new racemization activities, opens up the possibility for the rational design of new ACLRs with improved catalytic activity towards non-native substrates. This can ultimately lead to the development of improved biocatalysts that can meet the needs of industry.



Scheme 7 The suggested catalytic mechanism for the reaction catalyzed by ACLR. The two residues proposed as catalytic bases, Tyr₁₃₇ and D₂₁₀, are represented by the letter "X".

Cystathionine β -lyase (CBL)

Volatile sulfur compounds (VSCs) are key contributors to the flavor of a wide range of fermented foods, being of extreme relevance in cheese and fermented beverages, such as beer and wine.⁹⁸ These compounds are essential for the aroma of some wines, such as "grapefruit", "passion fruit", and "boxwood", having a great impact on the desired sensorial properties of the final product.⁹⁹ In cheese, compounds such as methanethiol, dimethyl disulfide, dimethyl trisulfide, and hydrogen sulfide have been linked to their characteristic flavor.¹⁰⁰ On the other hand, some VSCs might (above a certain concentration) induce a repulsive flavor, which has a detrimental effect on their quality.⁹⁸ Therefore, the formation of VSCs is a subject of great interest in the food industry, since fine control of their levels is required to maintain the highest quality of the manufactured products.

VSCs derive from sulfur-bearing precursors, such as methionine and cysteine, being products of their catabolism by bacteria. One enzyme that has been linked to the production of VSCs in food is the PLP-dependent enzyme, cystathionine β -lyase (CBL, EC 4.4.1.8).¹⁰¹ Commonly known as β -cystathionase, CBL is the second enzyme involved in the transsulfuration pathway for L-methionine biosynthesis, a pathway present in several bacteria and plants, but absent in humans. For this reason, CBL has been considered an attractive antibiotic target, with less chance of harmful and unwanted side effects.^{102,103} As mentioned, CBL, belonging to the γ -family of PLP-dependent enzymes, needs PLP as a co-factor.¹⁰⁴ Its major function is to catalyze the cleavage of the carbon-sulfur bond of L-cystathionine, releasing L-homocysteine, pyruvate and ammonia. However, CBL is not only involved in the synthesis of methionine, but also in its catabolism, as detected in the lactic acid bacteria *Lactococcus lactis* by the conversion of methionine into methanethiol.¹⁰⁵ Moreover, it was recently demonstrated that MetC, a CBL from *Escherichia coli*, besides its role in the L-methionine metabolism, is also able to produce non-canonical D-amino acids and is involved in the D- and L-serine metabolism.¹⁰⁶

For these reasons, CBL is considered as an enzyme with great potential for the chemical and pharmaceutical industries. Nevertheless, more knowledge about its structure and catalytic mechanism is required.

CBL is a homotetrameric enzyme with a PLP fold-type I and a chiral dihedral symmetry D₂ (Fig. 7). Each monomer of the enzyme is composed of three distinct domains. The N-terminal domain (residues 1 to 60), formed by three α -helices and one β -strand, is involved in tetramer formation and constitutes some of the active sites of the adjacent subunit. The second domain (residues 61 to 256) is formed by an α/β -structure and is the place where PLP is anchored to the enzyme. Lastly, the C-terminal (residues 257 to 395) consists of four helices, exposed to the solvent, and four-stranded antiparallel β -sheets that are located in the central part of this domain.¹⁰⁷

According to the crystallographic structure of CBL from *Escherichia coli* (PDB code 1CL1), obtained by Clausen and

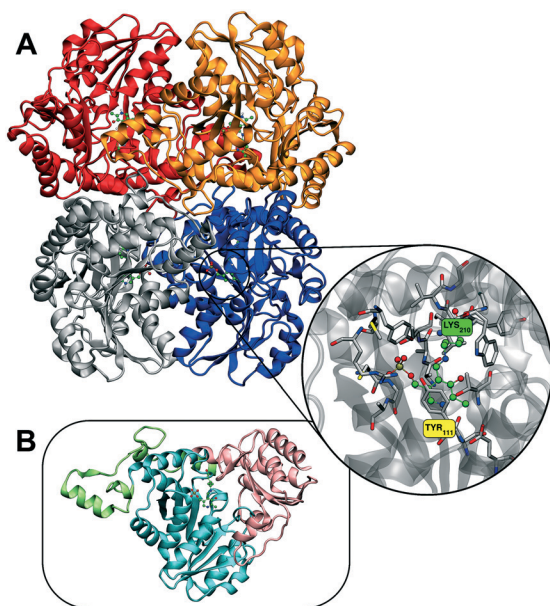
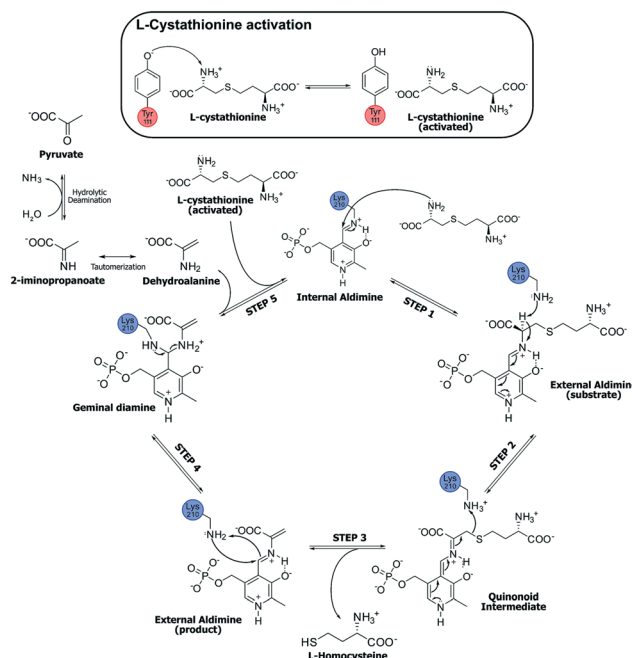


Fig. 7 (A) 3D structure of CBL from *Escherichia coli* (PDB code: 1CL1). In the bottom, the active site of the enzyme is presented with the internal aldimine represented in green. (B) The three domains of this enzyme present in each monomer are represented in different colours: N-terminal domain (light green), second domain (cyan), and C-terminal domain (pink).

his co-workers, there are two residues that can play a crucial role in the catalytic mechanism of this enzyme (Scheme 8): Lys210, and Tyr111. Lys210, the PLP-binding residue, is also responsible for the proton transfer between C_{α} and S_{γ} . Tyr111, on the other hand, is the residue responsible for substrate activation, being conserved in all cystathionine utilizing enzymes. It exists in the active site in a phenolate form, being responsible for the removal of a proton from the α -amino group of the substrate, cystathionine. The deprotonated amino group of the activated substrate can then perform a nucleophilic attack, displacing the lysine bound to PLP and forming a Schiff base designated external aldimine. The free lysine can now abstract the proton from the C_{α} to form a quinonoid intermediate. Subsequently, the protonation of S_{γ} leads to the cleavage of the $C\beta$ – S_{γ} bond. Then, the external aldimine is displaced by the nucleophilic attack of the lysine, regenerating the catalytic active internal aldimine and releasing the unstable enamine, dehydroalanine. This compound then tautomerizes into imine 2-iminopropanoate which undergoes hydrolytic deamination to form pyruvate and ammonia, a step that is unlikely to occur in the active site of the enzyme.¹⁰⁷

Enzymes involved in the formation of VSCs are of great interest for the food industry to produce compounds responsible for the flavor of a wide range of fermented food and beverages, such as cheese, wine, and beer. Cystathionine β -lyase (CBL) is a PLP-dependent enzyme involved in the production of VSCs, such as methanethiol and dimethyl disulfide. It is also able to synthesize D-amino acids, a characteristic that is of great interest to the chemical and



Scheme 8 Representation of the reaction catalyzed by cystathionine β -lyase.

pharmaceutical industries. However, the role played by CBL in the generation of these compounds is not completely understood. In order for that to happen, a thorough characterization at a molecular level of the enzyme needs to be done. Only then can CBL be optimized for industrial purposes.

Conclusions and future perspectives

As illustrated in this review, the PLP-dependent enzymes ω -transaminases, lysine decarboxylase, threonine aldolase, L-tyrosine phenol-lyase, α -amino- ϵ -caprolactam racemases, and cystathionine β -lyase are attractive targets for enzyme engineering towards their application in an IB framework. While the common element here present (*i.e.* PLP) plays an important role in all these enzymes, the catalytic mechanisms adopted have a variety of differences and subtleties that render the application of atomic-level computational enzymatic methods necessary to fully understand the catalytic determinants.

ω -Transaminases and threonine aldolase are good examples of enzymes for IB. There is already a good body of knowledge concerning their specific catalytic mechanisms and the role of the different amino acid residues in catalysis that was able to guide the development of a new generation of bio-engineered enzymes with increased reaction rates and/or specificity. The same can also be said about ATA 117 which is employed in the catalysis of the last step of the synthesis of the antidiabetic compound sitagliptin and was able to generate around 6 billion U.S. dollars of revenue to Merck only in 2017. This success was accomplished due to the bioengineering of the enzyme that has allowed improvement of the catalytic process in almost 13 percent while decreasing the total

manufacturing costs. These kinds of studies have granted the laureate Frances H. Arnold the Nobel Prize in Chemistry 2018, “for the directed evolution of enzymes”. The expression of these enzymes in bacterial strains is also something to recognize since without this goal, it would be more difficult and expensive to obtain desirable amounts of enzyme.⁵⁴

Although the success of these enzymes in IB envisages the potential of bioengineered enzymes for industrial applications, the same level of knowledge is not available for other enzymes. In fact, as it was reviewed in this article, many enzymes have the potential to be used in industrial applications. However, there is still a lack of knowledge that precludes their current application. To date, there are more questions about how they work than answers, representing the need for more investment in this area. However, there is no doubt that enzymes will be the future of IB and will be the secret behind the success of many industries.

Hence, despite the promising results obtained, which demonstrate the potential of these enzymes, more studies are required to fully master the activity of PLP-dependent enzymes for biocatalysis.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank FCT (Fundação para a Ciência e a Tecnologia) for Contract Grants IF/00052/2014 and for PhD Grants SFRH/BD/136746/2018, SFRH/BD/136594/2018 and UID/Multi/04378/2019.

References

- 1 K. E. Cassimjee, B. Manta and F. Himo, *Org. Biomol. Chem.*, 2015, **13**, 8453–8464.
- 2 R. A. Sheldon and P. C. Pereira, *Chem. Soc. Rev.*, 2017, **46**, 2678–2691.
- 3 R. A. Sheldon and J. M. Woodley, *Chem. Rev.*, 2018, **118**, 801–838.
- 4 R. A. Sheldon and D. Brady, *Chem. Commun.*, 2018, **54**, 6088–6104.
- 5 N. M. F. S. A. Cerqueira, P. A. Fernandes and M. J. Ramos, *ChemPhysChem*, 2018, **19**, 669–689.
- 6 T. Davids, M. Schmidt, D. Bottcher and U. T. Bornscheuer, *Curr. Opin. Chem. Biol.*, 2013, **17**, 215–220.
- 7 S. F. Sousa, A. J. M. Ribeiro, R. P. P. Neves, N. F. Bras, N. M. F. S. A. Cerqueira, P. A. Fernandes and M. J. Ramos, *WIREs Comput. Mol. Sci.*, 2017, **7**, e1281.
- 8 S. F. Sousa, J. F. Sousa, A. C. Barbosa, C. E. Ferreira, R. P. Neves, A. J. Ribeiro, P. A. Fernandes and M. J. Ramos, *J. Phys. Chem. A*, 2016, **120**, 5300–5306.
- 9 P. Ferreira, S. F. Sousa, P. A. Fernandes and M. J. Ramos, *Chemistry*, 2017, **23**, 17231–17241.
- 10 Y. Li and P. C. Cirino, *Biotechnol. Bioeng.*, 2014, **111**, 1273–1287.
- 11 H. C. Dunathan, *Proc. Natl. Acad. Sci. U. S. A.*, 1966, **55**, 712–716.
- 12 H. S. Fernandes, M. J. Ramos and N. M. F. S. A. Cerqueira, *Chemistry*, 2017, **23**, 9162–9173.
- 13 H. S. Fernandes, M. J. Ramos and N. M. F. S. A. Cerqueira, *ACS Catal.*, 2018, **8**, 10096–10110.
- 14 N. V. Grishin, M. A. Phillips and E. J. Goldsmith, *Protein Sci.*, 1995, **4**, 1291–1304.
- 15 R. Percudani and A. Peracchi, *BMC Bioinf.*, 2009, **10**, 273.
- 16 J. Liang, Q. Han, Y. Tan, H. Ding and J. Li, *Front. Mol. Biosci.*, 2019, **6**, 4.
- 17 H.-P.-T. Ngo, N. M. F. S. A. Cerqueira, J.-K. Kim, M.-K. Hong, P. A. Fernandes, M. J. Ramos and L.-W. Kang, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2014, **70**, 596–606.
- 18 E. F. Oliveira, N. M. F. S. A. Cerqueira, P. A. Fernandes and M. J. Ramos, *J. Am. Chem. Soc.*, 2011, **133**, 15496–15505.
- 19 N. M. F. S. A. Cerqueira, P. A. Fernandes and M. J. Ramos, *J. Chem. Theory Comput.*, 2011, **7**, 1356–1368.
- 20 D. Ghislieri and N. J. Turner, *Top. Catal.*, 2014, **57**, 284–300.
- 21 M. D. Patil, G. Grogan, A. Bommaris and H. Yun, *Catalysts*, 2018, **8**, 254.
- 22 A. Gomm and E. O'Reilly, *Curr. Opin. Chem. Biol.*, 2018, **43**, 106–112.
- 23 E. E. Ferrandi and D. Monti, *World J. Microbiol. Biotechnol.*, 2018, **34**, 13.
- 24 S. A. Kelly, S. Pohle, S. Wharry, S. Mix, C. C. R. Allen, T. S. Moody and B. F. Gilmore, *Chem. Rev.*, 2018, **118**, 349–367.
- 25 M. Höhne and U. T. Bornscheuer, *ChemCatChem*, 2009, **1**, 42–51.
- 26 J. Kim, D. Kyung, H. Yun, B. K. Cho, J. H. Seo, M. Cha and B. G. Kim, *Appl. Environ. Microbiol.*, 2007, **73**, 1772–1782.
- 27 F. Steffen-Munsberg, C. Vickers, H. Kohls, H. Land, H. Mallin, A. Nobili, L. Skalden, T. van den Bergh, H. J. Joosten, P. Berglund, M. Höhne and U. T. Bornscheuer, *Biotechnol. Adv.*, 2015, **33**, 566–604.
- 28 T. Pavkov-Keller, G. A. Strohmeier, M. Diepold, W. Peeters, N. Smeets, M. Schürmann, K. Gruber, H. Schwab and K. Steiner, *Sci. Rep.*, 2016, **6**, 38183.
- 29 A. Łyskowski, C. Gruber, G. Steinkellner, M. Schürmann, H. Schwab, K. Gruber and K. Steiner, *PLoS One*, 2014, **9**, e87350.
- 30 R. L. Hanson, B. L. Davis, Y. Chen, S. L. Goldberg, W. L. Parker, T. P. Tully, M. A. Montana and R. N. Patel, *Adv. Synth. Catal.*, 2008, **350**, 1367–1375.
- 31 S. Gao, Y. Su, L. Zhao, G. Li and G. Zheng, *Process Biochem.*, 2017, **63**, 130–136.
- 32 M. Höhne, S. Schätzle, H. Jochens, K. Robins and U. T. Bornscheuer, *Nat. Chem. Biol.*, 2010, **6**, 807–813.
- 33 C. Iglesias, P. Panizza and S. Rodríguez Giordano, *Appl. Microbiol. Biotechnol.*, 2017, **101**, 5677–5687.
- 34 A. Iwasaki, K. Matsumoto, J. Hasegawa and Y. Yasohara, *Appl. Microbiol. Biotechnol.*, 2012, **93**, 1563–1573.
- 35 J. Jiang, X. Chen, D. Zhang, Q. Wu and D. Zhu, *Appl. Microbiol. Biotechnol.*, 2015, **99**, 2613–2621.
- 36 F. G. Mutti, C. S. Fuchs, D. Pressnitz, J. H. Sattler and W. Kroutil, *Adv. Synth. Catal.*, 2011, **353**, 3227–3233.

- 37 S. Schätzle, F. Steffen-Munsberg, A. Thontowi, M. Höhne, K. Robins and U. T. Bornscheuer, *Adv. Synth. Catal.*, 2011, **353**, 2439–2445.
- 38 M. Thomsen, L. Skalden, G. J. Palm, M. Höhne, U. T. Bornscheuer and W. Hinrichs, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2014, **70**, 1086–1093.
- 39 S. W. Han, E. S. Park, J. Y. Dong and J. S. Shin, *Appl. Environ. Microbiol.*, 2015, **81**, 6994–7002.
- 40 I. V. Pavlidis, M. S. Weiß, M. Genz, P. Spurr, S. P. Hanlon, B. Wirz, H. Iding and U. T. Bornscheuer, *Nat. Chem.*, 2016, **8**, 1076–1082.
- 41 C. K. Savile, J. M. Janey, E. C. Mundorff, J. C. Moore, S. Tam, W. R. Jarvis, J. C. Colbeck, A. Krebber, F. J. Fleitz, J. Brands, P. N. Devine, G. W. Huisman and G. J. Hughes, *Science*, 2010, **329**, 305–309.
- 42 M. S. Weiß, I. V. Pavlidis, P. Spurr, S. P. Hanlon, B. Wirz, H. Iding and U. T. Bornscheuer, *Org. Biomol. Chem.*, 2016, **14**, 10249–10254.
- 43 M. S. Weiß, I. V. Pavlidis, P. Spurr, S. P. Hanlon, B. Wirz, H. Iding and U. T. Bornscheuer, *ChemBioChem*, 2017, **18**, 1022–1026.
- 44 A. A. Desai, *Angew. Chem., Int. Ed.*, 2011, **50**, 1974–1976.
- 45 S. Lutz, *Science*, 2010, **329**, 285–287.
- 46 L. Martinez-Montero, J. H. Schrittwieser and W. Kroutil, *Top. Catal.*, 2018, DOI: 10.1007/s11244-018-1054-7.
- 47 D. F. A. R. Dourado, S. Pohle, A. T. P. Carvalho, D. S. Dheeman, J. M. Caswell, T. Skvortsov, I. Miskelly, R. T. Brown, D. J. Quinn, C. C. R. Allen, L. Kulakov, M. Huang and T. S. Moody, *ACS Catal.*, 2016, **6**, 7749–7759.
- 48 F. Steffen-Munsberg, C. Vickers, A. Thontowi, S. Schätzle, T. Meinhardt, M. SvedendahlHumble, H. Land, P. Berglund, U. T. Bornscheuer and M. Höhne, *ChemCatChem*, 2013, **5**, 154–157.
- 49 L. J. Guan, J. Ohtsuka, M. Okai, T. Miyakawa, T. Mase, Y. Zhi, F. Hou, N. Ito, A. Iwasaki, Y. Yasohara and M. Tanokura, *Sci. Rep.*, 2015, **5**, 10753.
- 50 P. Berglund, M. S. Humble and C. Branneby, in *Comprehensive Chirality*, ed. E. M. Carreira and H. B. T. Yamamoto, Elsevier, Amsterdam, 2012, vol. 7, pp. 390–401.
- 51 K. E. Cassimjee, M. S. Humble, V. Miceli, C. G. Colomina and P. Berglund, *ACS Catal.*, 2011, **1**, 1051–1055.
- 52 M. Svedendahl, C. Branneby, L. Lindberg and P. Berglund, *ChemCatChem*, 2010, **2**, 976–980.
- 53 W. Ma, K. Chen, Y. Li, N. Hao, X. Wang and P. Ouyang, *Engineering*, 2017, **3**, 308–317.
- 54 H. T. Kim, K.-A. Baritugo, Y. H. Oh, S. M. Hyun, T. U. Khang, K. H. Kang, S. H. Jung, B. K. Song, K. Park, I.-K. Kim, M. O. Lee, Y. Kam, Y. T. Hwang, S. J. Park and J. C. Joo, *ACS Sustainable Chem. Eng.*, 2018, **6**, 5296–5305.
- 55 S. Bunsupa, K. Katayama, E. Ikeura, A. Oikawa, K. Toyooka, K. Saito and M. Yamazaki, *Plant Cell*, 2012, **24**, 1202–1216.
- 56 S. H. Park, F. Soetyono and H. K. Kim, *J. Microbiol. Biotechnol.*, 2017, **27**, 289–296.
- 57 J. Schneider and V. F. Wendisch, *Appl. Microbiol. Biotechnol.*, 2011, **91**, 17.
- 58 E. Kandiah, D. Carriel, J. Perard, H. Malet, M. Bacia, K. Liu, S. W. S. Chan, W. A. Houry, S. Ollagnier de Choudens, S. Elsen and I. Gutsche, *Sci. Rep.*, 2016, **6**, 24601.
- 59 S. Kind, S. Kreye and C. Wittmann, *Metab. Eng.*, 2011, **13**, 617–627.
- 60 E. F. Gale and H. M. Epps, *Biochem. J.*, 1944, **38**, 232–242.
- 61 K. Fesko, *Appl. Microbiol. Biotechnol.*, 2016, **100**, 2579–2590.
- 62 H. J. Gwon and S. H. Baik, *Biotechnol. Lett.*, 2010, **32**, 143–149.
- 63 T. A. Scott, D. Heine, Z. Qin and B. Wilkinson, *Nat. Commun.*, 2017, **8**, 15935.
- 64 Q. Chen, X. Chen, Y. Cui, J. Ren, W. Lu, J. Feng, Q. Wu and D. Zhu, *Catal. Sci. Technol.*, 2017, **7**, 5964–5973.
- 65 S. E. Franz and J. D. Stewart, *Adv. Appl. Microbiol.*, 2014, **88**, 57–101.
- 66 C. L. Kielkopf and S. K. Burley, *Biochemistry*, 2002, **41**, 11711–11720.
- 67 M. K. Uhl, G. Oberdorfer, G. Steinkellner, L. Riegler-Berket, D. Mink, F. van Assema, M. Schurmann and K. Gruber, *PLoS One*, 2015, **10**, e0124056.
- 68 S. H. Balk, H. Yoshioka, H. Yukawa and S. Harayama, *J. Microbiol. Biotechnol.*, 2007, **17**, 721–727.
- 69 M. L. di Salvo, S. G. Remesh, M. Vivoli, M. S. Ghatge, A. Paiardini, S. D'Aguzzo, M. K. Safo and R. Contestabile, *Rev. Geophys.*, 2014, **281**, 129–145.
- 70 T. Kimura, V. P. Vassilev, G.-J. Shen and C.-H. Wong, *J. Am. Chem. Soc.*, 1997, **119**, 11734–11742.
- 71 H. M. Qin, F. L. Imai, T. Miyakawa, M. Kataoka, N. Kitamura, N. Urano, K. Mori, H. Kawabata, M. Okai, J. Ohtsuka, F. Hou, K. Nagata, S. Shimizu and M. Tanokura, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2014, **70**, 1695–1703.
- 72 L. Wieteska, M. Ionov, J. Szemraj, C. Feller, A. Kolinski and D. Gront, *J. Biotechnol.*, 2015, **199**, 69–76.
- 73 S. Kurjatschij, M. Katzberg and M. Bertau, *J. Mol. Catal. B: Enzym.*, 2014, **103**, 3–9.
- 74 J. D. Tibhe, H. Fu, T. Noel, Q. Wang, J. Meuldijk and V. Hessel, *Beilstein J. Org. Chem.*, 2013, **9**, 2168–2179.
- 75 S. T. Gurme, S. N. Surwase, S. A. Patil, S. B. Jadhav and J. P. Jadhav, *Indian J. Microbiol.*, 2013, **53**, 194–198.
- 76 S. Chattopadhyay, S. K. Datta and S. B. Mahato, *Plant Cell Rep.*, 1994, **13**, 519–522.
- 77 W. S. Knowles, *J. Chem. Educ.*, 1986, **63**, 222–225.
- 78 Q. Do, G. T. Nguyen and R. S. Phillips, *Amino Acids*, 2016, **48**, 2243–2251.
- 79 T. Li and X. Li, *Green Chem.*, 2014, **16**, 4241–4256.
- 80 R.-C. Zheng, X.-L. Tang, H. Suo, L.-L. Feng, X. Liu, J. Yang and Y.-G. Zheng, *Enzyme Microb. Technol.*, 2018, **112**, 88–93.
- 81 A. A. Antson, T. V. Demidkina, P. Gollnick, Z. Dauter, R. L. von Tersch, J. Long, S. N. Berezhnoy, R. S. Phillips, E. H. Harutyunyan and K. S. Wilson, *Biochemistry*, 1993, **32**, 4195–4206.
- 82 D. Milić, T. V. Demidkina, N. G. Faleev, R. S. Phillips, D. Matković-Čalogović and A. A. Antson, *J. Am. Chem. Soc.*, 2011, **133**, 16468–16476.

- 83 B. Sundararaju, A. A. Antson, R. S. Phillips, T. V. Demidkina, M. V. Barbolina, P. Gollnick, G. G. Dodson and K. S. Wilson, *Biochemistry*, 1997, **36**, 6502–6510.
- 84 D. Milic, T. V. Demidkina, N. G. Faleev, D. Matkovic-Calogovic and A. A. Antson, *J. Biol. Chem.*, 2008, **283**, 29206–29214.
- 85 N. Stabler, T. Oikawa, M. Bott and L. Eggeling, *J. Bacteriol.*, 2011, **193**, 1702–1709.
- 86 Y. Asano and S. Yamaguchi, *J. Am. Chem. Soc.*, 2005, **127**, 7696–7697.
- 87 Y. Asano and S. Yamaguchi, *J. Mol. Catal. B: Enzym.*, 2005, **36**, 22–29.
- 88 S. Yamaguchi, H. Komeda and Y. Asano, *Appl. Environ. Microbiol.*, 2007, **73**, 5370–5373.
- 89 S. A. Ahmed, N. Esaki, H. Tanaka and K. Soda, *Agric. Biol. Chem.*, 1983, **47**, 1887–1893.
- 90 T. Fukumura, *Agric. Biol. Chem.*, 1977, **41**, 1327–1330.
- 91 S. A. Ahmed, N. Esaki, H. Tanaka and K. Soda, *FEBS Lett.*, 1984, **174**, 76–79.
- 92 A. Frese, S. V. Barrass, P. W. Sutton, J. P. Adams and G. Grogan, *ChemBioChem*, 2018, **19**, 1711–1715.
- 93 S. Okazaki, A. Suzuki, T. Mizushima, T. Kawano, H. Komeda, Y. Asano and T. Yamane, *Biochemistry*, 2009, **48**, 941–950.
- 94 S. A. Ahmed, N. Esaki, H. Tanaka and K. Soda, *Biochemistry*, 1986, **25**, 385–388.
- 95 R. Awad, P. Gans and J.-B. Reiser, *Biochimie*, 2017, **137**, 165–173.
- 96 V. W. C. Soo, Y. Yosaatmadja, C. J. Squire and W. M. Patrick, *J. Biol. Chem.*, 2016, **291**, 19873–19887.
- 97 A. Frese, P. W. Sutton, J. P. Turkenburg and G. Grogan, *ACS Catal.*, 2017, **7**, 1045–1048.
- 98 S. Landaud, S. Helinck and P. Bonnarne, *Appl. Microbiol. Biotechnol.*, 2008, **77**, 1191–1205.
- 99 S. Holt, A. G. Cordente, S. J. Williams, D. L. Capone, W. Jitjaroen, I. R. Menz, C. Curtin and P. A. Anderson, *Appl. Environ. Microbiol.*, 2011, **77**, 3626–3632.
- 100 K. E. Seefeldt and B. C. Weimer, *J. Dairy Sci.*, 2000, **83**, 2740–2746.
- 101 W. J. Lee, D. S. Banavara, J. E. Hughes, J. K. Christiansen, J. L. Steele, J. R. Broadbent and S. A. Rankin, *Appl. Environ. Microbiol.*, 2007, **73**, 3034–3039.
- 102 L. J. Ejim, J. E. Blanchard, K. P. Koteva, R. Sumerfield, N. H. Elowe, J. D. Chechetto, E. D. Brown, M. S. Junop and G. D. Wright, *J. Med. Chem.*, 2007, **50**, 755–764.
- 103 A. Astegno, A. Giorgetti, A. Allegrini, B. Cellini and P. Dominici, *BioMed Res. Int.*, 2013, **2013**, 1–13.
- 104 F. W. Alexander, E. Sandmeier, P. K. Mehta and P. Christen, *Eur. J. Biochem.*, 1994, **219**, 953–960.
- 105 A. C. Alting, W. J. M. Engels, S. Vanschalkwijk and F. A. Exterkate, *Appl. Environ. Microbiol.*, 1995, **61**, 4037–4042.
- 106 T. Miyamoto, M. Katane, Y. Saitoh, M. Sekine and H. Homma, *Biochem. J.*, 2018, **475**, 1397–1410.
- 107 T. Clausen, R. Huber, B. Laber, H.-D. Pohlenz and A. Messerschmidt, *J. Mol. Biol.*, 1996, **262**, 202–224.

FACULDADE DE **MEDICINA**

