U.PORTO
INSTITUTO DE CIÊNCIAS
BIOMÉDICAS ABEL SALAZAR
UNIVERSIDADE DO PORTO

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

MSc

2.º
CICLO

FCUP
2020

U.PORTO

Identification of New Drugs Against Biofilm Formation and Development

Fábio António Gouveia Ferreira Martins

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

U.PORTO
INSTITUTO DE CIÊNCIAS
BIOMÉDICAS ABEL SALAZAR
UNIVERSIDADE DO PORTO

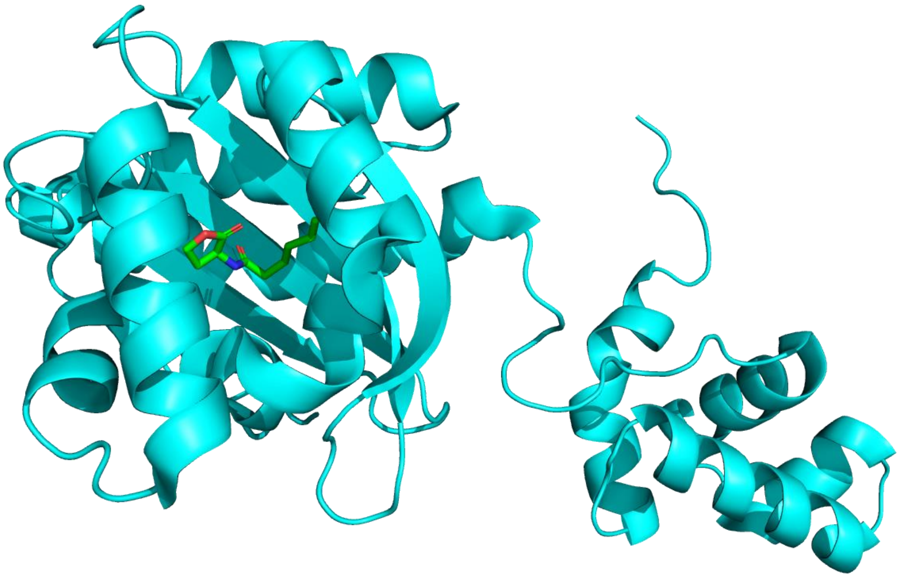
Identification of New Drugs Against Biofilm Formation and Development

Fábio António Gouveia Ferreira Martins

Dissertação de Mestrado apresentada ao Instituto de Ciências Biomédicas Abel Salazar e à Faculdade de Ciências da Universidade do Porto em Bioquímica

2020

MSc



Identification of New Drugs Against Biofilm Formation and Development

Fábio António Gouveia Ferreira Martins
Mestrado em Bioquímica
Faculdade de Ciências da Universidade do Porto
Instituto de Ciências Biomédicas Abel Salazar
2020

Orientador
André Alberto de Sousa Melo
Professor Auxiliar, FCUP
Coorientador
Sérgio Filipe Maia de Sousa
Investigador Auxiliar, UCIBIO/REQUIMTE - BioSIM

U.PORTO

INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
UNIVERSIDADE DO PORTO

U.PORTO

FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ______/______/_________

MSC

Agradecimentos

Com o finalizar deste ano de trabalho, gostaria de agradecer a todos que tornaram esta dissertação possível.

Em primeiro lugar, ao meu orientador, o Professor André Melo, não só pela oportunidade de trabalhar com ele, mas também pela disponibilidade e aprendizagem que me ofereceu ao longo deste ano.

Ao meu co-orientador, o Professor Sérgio Sousa, por toda a ajuda, orientação e apoio que diariamente me ofereceu.

Ao BioSIM, por me fazerem sentir bem vindo, por toda a conversa, e claro, por estarem sempre disponíveis para ajudar quando tinha qualquer problema. A todos, obrigado.

Aos meus amigos por me distraírem sempre que preciso, e pela amizade que me oferecem.

E por fim, à minha família, por todo o apoio, paciência e encorajamento que me deram ao longo de toda a minha vida e pelos sacrifícios que fazem diariamente que permitiram a realização dos meus estudos.

Abstract

Bacterial biofilms are aggregates of microorganisms anchored to a surface and embedded in a self-produced matrix of extracellular polymeric substances. Thanks to the properties of the matrix and intercellular interactions between the bacteria within, the biofilm becomes increasingly sophisticated with the bacteria having different behaviours when compared to their planktonic counterparts. Microorganisms in biofilms have several advantages including an increased antibiotic resistance, elevated levels of lateral gene transfer, higher stress resistance and subversion of host defence mechanisms. Biofilm formation is a complex procedure, involving four phases: attachment to a surface, sessile growth phase, biofilm maturation and detachment when the biofilm is no longer beneficial.

Cell-to-cell communication, or quorum sensing, is an important process during biofilm maturation, in which cells communicate using auto-inducer signals. It is also reported that various virulence factors are regulated by quorum sensing.

Biofilm infections have been recognized as a serious threat to our society. However, although our knowledge about biofilms is increasing, the ability to control biofilm formation and to treat biofilm infections remains insufficient. Therefore, the aim of this work is to help find new drugs against this issue. Because quorum sensing has specific protein targets, it is possible to design inhibitors to block the formation of these structures.

The process of drug discovery and development is a long and expensive process. Over the years, the usage of computer-aided drug design as a preliminary stage of drug design has increased in order to make the entire process more cost-efficient and minimize failures. CADD is defined by IUPAC as “all computer assisted techniques used to discover, design and optimize compounds with desired structures and properties”. These techniques are used to screen large compound libraries for promising molecules to be tested experimentally, optimize lead compounds or design new drugs against a specific target.

Protein-Ligand molecular docking is a computational tool which predicts the binding pose and affinity of a ligand to a specific receptor or enzyme. During a virtual screening procedure, thousands or even millions of molecules are docked into a particular target and scored, giving an indication to which molecules are more probable to be active. This reduces the number of molecules that have to be experimentally tested.

This work is focused in discovering new promising compounds against CviR, the quorum sensing receptor from *Chromobacterium violaceum*. This bacterium is an opportunistic pathogen used as a model organism for QS research. Before the actual virtual screening, it was necessary to create an optimized protocol for this target. The optimization of the protocol involved several steps. The first step was to download the six available pdb structures of CviR. In order to see how accurately the molecular docking programs, reproduce experimentally known complexes, multiple redocking procedures of the original ligand into their native PDB structures were performed. The programs used for molecular docking were Autodock 4, Autodock Vina, GOLD and LeDock. To further evaluate both the programs and the available pdb structures, crossdocking of the original ligands from each structure into the other structures was executed. The ability to discriminate the active molecules within a large database was optimized by screening a library containing known active molecules and decoys. The optimized protocol was then applied to a ZINC/FDA Approved database and to the Mu.Ta.Lig Virtual Chemotheca, which resulted in a list of promising compounds for further studies.

Following the virtual screening stage of this work, Molecular dynamics simulations of the most promising molecules, in complex with CviR, were performed. Using the last 40 ns of simulation, MM/PBSA and MM/GBSA calculations were done in order to estimate the affinity of each molecule towards CviR. From these calculations, six compounds were found that have better or similar results than the native referebce ligand.

In this work, the development of a computational protocol for the virtual screening of large compound libraries and further analysis though molecular dynamics simulations and MM/PB(GB)SA calculations was reported. This protocol can now be applied to other databases, allowing the discovery of additional promising compounds which can be tested and validated experimentally.

Keywords: Bacteria, Biofilms, *Chromobacterium violaceum*, CviR, Quorum-Sensing, Virtual Screening, Molecular Docking, Molecular Dynamics Simulations, MM/PBSA, MM/GBSA.

Resumo

Biofilmes bacterianos são agregados de microrganismos, ancorados a uma superfície e envolvidos numa matriz extracelular, produzida pelas próprias bactérias, constituída por substâncias poliméricas. Devido às propriedades da matriz e a interações intercelulares entre as bactérias, o biofilme torna-se mais complexo e as bactérias dentro do biofilme adoptam diferentes comportamentos das suas homólogas planctónicas. Ao fazer parte de biofilmes, os microrganismos adquirem várias vantagens, incluindo uma maior resistência aos antibióticos, níveis elevados de transferência horizontal de genes; maior resistência ao stress e uma maior resiliência contra os mecanismos de defesa do hospedeiro. A formação de biofilmes é um processo complexo que envolve quatro fases: fixação à superfície, fase de crescimento fixa, maturação do biofilme e a separação quando o biofilme já não for benéfico.

A comunicação celular, ou *quorum sensing*, é um processo importante durante a maturação do biofilme, no qual as células comunicam usando sinais auto indutores. Vários fatores de virulência também são regulados via este mecanismo.

Infeções relacionadas com biofilmes são reconhecidas como uma séria ameaça. No entanto, apesar de o nosso conhecimento sobre biofilmes estar a aumentar, a nossa capacidade de controlar a formação de biofilmes e de tratar infeções por eles causadas é insuficiente. Assim, o objetivo deste trabalho é ajudar à procura de novos fármacos contra este problema. Uma vez que o processo de *quorum sensing* envolve alvos proteicos específicos, é possível encontrar inibidores que impeçam a formação destas estruturas.

O processo de procura e desenvolvimento de fármacos é um processo longo e dispendioso. Ao longo dos anos, na fase inicial do processo, tem vindo a aumentar o papel desempenhado pelo uso de desenho de fármacos assistido por computador. Usar métodos computacionais numa fase inicial do processo minimiza as falhas e os custos. O CADD (do inglês *computer-aided drug design)* é definido pela IUPAC como todas as técnicas computacionais usadas para descobrir, desenhar e otimizar moléculas com as estruturas e propriedades desejadas. Estas técnicas são usadas para testar grandes bases de dados em busca de moléculas promissoras para serem testadas experimentalmente, para otimizar compostos promissores já conhecidos ou para desenhar novos inibidores para um alvo específico.

Docking molecular proteína-ligando é um método computacional que prevê a pose e a afinidade entre um ligando e um recetor proteico específico. Durante um *virtual* *screening*, milhares ou até milhões de processos de *docking* são efetuados. Cada molécula recebe uma pontuação que prevê a afinidade desse ligando para o alvo proteico. Esta previsão de quais moléculas têm maior probabilidade de terem a afinidade desejada diminui o número de moléculas que serão testadas experimentalmente.

Este trabalho foi focado na procura de moléculas promissoras contra a CviR, o recetor de *quorum sensing* da *Chromobacterium violaceum.* Esta bactéria é um patógenico oportunista que é usado como modelo na investigação do *quorum sensing*. Antes de proceder ao *virtual screening,* é necessário desenvolver um protocolo otimizado para este alvo proteico. A otimização do protocolo envolveu vários passos. Em primeiro lugar, foram analisadas as seis estruturas da CviR que se encontravam disponíveis. De seguida procurou-se perceber se os programas de *docking* molecular são capazes de reproduzir os complexos experimentais. Os programas usados foram o Autodock 4, Autodock Vina, GOLD e o LeDock. De modo a avaliar não só os programas, mas também as estruturas, efetuou-se um *crossdocking* de cada ligando em todas as estruturas. Para melhorar a capacidade de reconhecer/discriminar moléculas ativas dentro de uma grande base de dados, o protocolo foi otimizado usando uma biblioteca contendo moléculas ativas contra a CviR (testadas experimentalmente), e moléculas inativas (*decoys).* O protocolo otimizado foi então aplicado à base de dados ZINC/FDA Approved e à Mu.Ta.Lig Virtual Chemotheca. Estes processos de *virtual screening* resultaram em vários compostos promissores.

Após a fase de *virtual screening*, foram efetuadas simulações de dinâmica molecular para as moléculas mais promissoras. Usando os últimos 40 ns de simulação, foram efetuados cálculos de MM/PBSA e MM/GBSA, que permitem estimar a afinidade de cada molécula para a CviR. Estes cálculos resultaram em seis compostos com afinidades iguais ou superiores ao ligando nativo.

Neste trabalho é reportado o desenvolvimento de um protocolo computacional para o *virtual screening* de grandes bases de dados, e de sucessiva análise via simulações de dinâmica molecular e cálculos de MM/PB(GB)SA. Este protocolo pode agora ser aplicado a outras bases de dados, permitindo a descoberta de mais moléculas promissoras que poderão ser testadas e validadas experimentalmente.

Table of Contents

Index of Figures

Index of Equations

Index of Tables

List of Abbreviations

AMP - Antimicrobial Peptides

ASP - Astex Statistical Potential

AHLs - Acyl Homoserine Lactones

AUC - Area Under the Curve

AI-2 - Autoinducer-2

AIPs - Autoinducing Peptides

BEDROC - Boltzmann Enhanced Discrimination of ROC

CADD - Computational Aided Drug Design

DUD-E - Database of Useful Decoys: Enhanced

EF - Enrichment Factor

EPS - Extracellular Polymeric Substance

FPR – False Positive Rate

GA – Genetic Algorithm

GOLD - Genetic Optimization for Ligand Docking

HTS – High Throughput Screening

QS - Quorum Sensing

QSAR - Quantitative Structure–Activity Relationship

QSP - Quorum Sensing Peptides

MD – Molecular Dynamics

MDR- Multidrug-resistant Bacteria

MM/GBSA - Molecular Mechanics/Generalized Born Surface Area

MM/PBSA - Molecular Mechanics/Poisson-Boltzmann Surface Area

NME - New Molecular Entities

PLP - Piecewise Linear Potential

RIE - Robust Initial Enhancement

RMSD - Root-mean-square deviation

ROC - Receiver Operating Characteristic

SASA - Solvent Accessible Surface Area

SMILES - Simplified Molecular Input Line Entry Specification

TG – Total Gain

TPR – True Positive Rate

VS- Virtual Screening

1. Introduction

1.1 Biofilms

1.1.1 Definition

Bacterial biofilms are aggregates of microorganisms in which cells are embedded in a self-produced matrix of extracellular polymeric substances, anchored to biotic or abiotic surfaces[1].

Most cells in biofilms experience cell-to-cell contact, and through intercellular interactions and the properties of the matrix, the biofilm becomes increasingly sophisticated in its activities. This makes the bacteria within these structures clearly different from their planktonic counterparts.[2,3] Microorganisms in biofilms have communal benefits such as increased antibiotic resistance, slow growth, differential gene expression, elevated levels of lateral gene transfer, stress resistance and subversion of host defence mechanisms[4–6].

1.1.2 Biofilm Formation

Biofilm formation occurs as a response to when the population density of unicellular bacteria reaches a certain threshold level. This process can be grouped into four phases: attachment, sessile growth phase, biofilm maturation and detachment[7].

In the first step microbial cells attach to a surface, which can be a tissue or an abiotic surface. Motile bacteria have been documented to present a competitive advantage, using flagella to overcome repulsive forces. In some bacterial species, chemotaxis also plays a role in directing attachment in response to nutrient composition[8]. The fimbriae, pilli and flagella give strength to the interaction between the bacteria and the surface of the attachment[9].

Surface contact triggers responses which lead to changes in gene expression. These changes up-regulate factors which favour sessility, such as those implicated in the formation of the extracellular polymeric matrix. After the attachment becomes stable, a process of multiplication and division starts, leading to the formation of many types of micro-communities that coordinate each other in multiple aspects[9].

The coordination between micro-communities is essential for the exchange of substrates, distribution of metabolic products and excretion of metabolic end-products[9]. Cell-to-cell communication is an important process during biofilm maturation, in which cells communicate using auto-inducer signals. When the required microbial density is attained, the secretion of signal molecules known as auto-inducers facilitates quorum sensing (QS). At this stage, certain proteins important for the formation of Extracellular Polymeric Substances (EPS), the main material in the biofilm's three-dimensional structure, are expressed. Interstitial voids are then produced within the matrix. These channels are filled with water and act as a circulatory system which is used to distribute nutrients and remove waste products from the micro-communities in the biofilm[9].

As biofilms mature, dispersal becomes an option. Other than passive dispersal, brought by shear stress, bacteria have evolved ways to react to the environment and decide whether it is still beneficial to reside within the biofilm or return to a planktonic lifestyle[8]. In this phase, microbial cells perform quick multiplication and dispersion in order to convert from sessile into motile form. The microbial communities produce saccharolytic enzymes, which are responsible for the lysis of the EPS matrix and subsequent detachment. In this phase, microbial cells upregulate the expression of protein related to flagella formation in order to let the bacteria move to a new site. This detachment of microbial cells and the transfer to a new location are important for the spreading of infections[10].

1.1.3 The EPS Matrix and Biofilm Architecture

Most of the biomass of the biofilm is comprised of EPS and not microbial cells. In most biofilms the microorganisms account for less than 10% of the dry mass whereas the matrix can account for more than 90 %. It consists of a conglomeration of different biopolymers, which form a scaffold for the biofilm architecture. This is responsible for the adhesion to surfaces and for cohesion in the biofilm[11].

The EPS composition can vary among biofilms, depending on the microorganisms present as well as the surrounding environment. Initially, EPS was thought to be composed only by extracellular polysaccharides. However it later became clear that this matrix also contains proteins, nucleic acids, lipids and other biopolymers[11]. Some of the polysaccharides are neutral or polyanionic, as is the case of the EPS of gram-negative bacteria. This allows the association of divalent cations such as calcium and magnesium which cross-link with the polymer strands and provide greater binding force to the biofilm. In the case of gram-positive bacteria however, the chemical composition of EPS may be primarily cationic. Extracellular bacterial structures such as flagella, pili and fimbriae can also stabilize the matrix[12].

EPS production is known to be affected by nutrient status of the growth medium. Excess of available carbon and limitation of nitrogen, phosphate or potassium promote EPS synthesis, as does slow bacterial growth[13].

The EPS has several important functions within the biofilm. The polysaccharides, proteins and DNA are involved in the adhesion in the initial steps of biofilm formation, immobilizing the cells within the biofilm. This keeps them in close proximity, allowing for intense interactions and the formation of the synergistic micro-comunitites[11]. Due to the existence of extracellular enzymes, an external digestive system is generated. This matrix also acts as a recycling centre by keeping the components of lysed cells available. This includes DNA, which acts as a reserve of genes for horizontal gene transfer. It is also important for the retention of water, leading to their tolerance in water-deficient environments, and it can be a nutrient source, providing a source of carbon, nitrogen, and phosphorus containing compounds. This matrix also acts as a protective barrier, making the biofilm resistant to nonspecific and specific host defences during infections, and to various antimicrobial agents[11].

Biofilms are very heterogeneous containing microcolonies of bacterial cells encased in an EPS matrix and separated from other microcolonies by interstitial voids[14]. This heterogeneity happens not only on mixed biofilms but also for pure culture biofilms such as the ones which are common on medical devices and those associated with infectious disease[14]. The organisms composing the biofilm may also have an effect on the structure, with the biofilm thickness affected by the number of component organisms[15].

The structure may also be influenced by particles from the host or environment. For example, erythrocytes and fibrin may accumulate in biofilms on heart valves. The fibrin will protect the organisms in the biofilm from the host's leukocytes, leading to infective endocarditis[14].

The EPS matrix can also dynamically modulate chemical and nutrient gradients and define pathogenic environments. These effects contribute to key virulence attributes, including recalcitrance. Thus, targeting the EPS matrix may be an effective strategy to remove biofilms, disaggregate bacteria and disrupt the pathogenic environment[16].

1.1.4 Resistance and Persister Cells

The terms persistence, resistance and tolerance are often confused when talking about the inability of antibiotics to kill or inhibit the growth of bacteria within a biofilm[17]. Resistance usually has a genetic basis and can be acquired though point mutation or horizontal gene transfer. Tolerance is better used when antibiotic-susceptible strains require much higher concentrations to obtain similar effects to the ones observed on planktonic cells. This tolerance can be lost when biofilms disperse into single cells. That way, dispersal strategies can be used as an adjuvant to the antibiotics. Persistence is used to describe a prolonged biofilm infection that remains after treatment[18].

Persistence should also not be confused with persister cells. These cells are in a state of dormancy, a state in which cells are metabolically inactive. This phenotype was first described in 1942 by Hobby et al. who found that 1% of *Staphylococcus aureus* cells were not killed by penicillin and became persister cells[19]. In 1944, Bigger discovered that one in a million *Staphylococcus pyogenes* cells were not killed by penicillin, and that, unlike resistant cells, these did not suffer any genetic change[20].

Expression profiling of RNA from persister cells revealed the downregulation of transcription genes involved in energy production and non-essential functions, which is consistent with the idea that these cells are in a state of dormancy[21].

Persister cells are believed to be responsible for the persitence of chronic biofilm infections. In fact, after antibiotics kill the majority of cells, persisters remain and repopulate the biofilms after the level of antibiotic drops. These cells are less susceptible to antibiotics because they are not undergoing metabolic activities that antibiotics can disturb. However, the resistant cells, arise from genetic changes that stop the antibiotic activity. This means that resistant cells continue to grow when antibiotics are present, unlike persister cells that are dormant and do not grow[21].

1.1.5 Clinical Significance

Several clinically important pathogenic bacteria such as cystic fibrosis associated *Pseudomonas aeruginosa,* urinary and catheter-associated *Proteus mirabilis,* lower respiratory tract and surgical sites associates *Staphylococcus aureus,* pneumonia causing *Haemophilus influenza,* and many others, cause infection through biofilms formation[7]. This can have devastating consequences, because as previous mentioned, microbes that reside in biofilms may not be eliminated by traditional antibiotics because of metabolic dormancy or molecular resistance mechanisms. The extracellular matrix also has an important role in conferring tolerance to biofilms[22].

The overall burden of biofilm infections is significant, and it has been recognized as a serious threat to our society within the past 20 years. The U.S. National Institutes of Health estimates that 80% of all bacterial infections occurring in the human body are biofilm related[23]. In the United States an estimated 17 million new biofilm infections occur each year, which can result in up to 550000 fatalities per year[24]. Among the over 60000 cystic fibrosis patients in developed western countries, nearly 80 % will develop a chronic biofilm lung infection. For the 1-2 % of western population with chronic wound infections, 60 % of these involve biofilms[25]. The development of biofilm on the surface of endotracheal tubes is related to the development of ventilator-associated pneumonia, which occurs in 9-27 % of all intubated patients[26]. On patients with indwelling urinary catheters, the rates are even higher with over 50 % of inserted catheters becoming colonized within the first 14 days of insertion[27].

Although our knowledge about biofilms is increasing, the control of biofilm formation and the treatment for existing biofilms remains tenuous, with few new therapeutic options currently available clinically[18].

1.1.6. Current therapeutic approaches

Because of the recalcitrance and the consequences of a persistent biofilm infection, the treatment may include old, last-resort antibiotics such as colistin[28]. This issue is worsened by the fact that multi-drug resistant bacteria are becoming more common which also increases the usage of last-resort antibiotics[24]. Other strategies are based on cancer treatment, including a high and sustained antimicrobial chemotherapy[29]. Another approach used for intravenous catheter infections is lock therapy[30]. In general, antibiotic lock solutions combine highly concentrated antibiotic with an anticoagulant to allow for local instillation into the catheter lumen. The solution is locked while the catheter is not in use to prevent colonization or sterilize a previously infected catheter[31]. However, since killing does not necessarily eradicate the biofilm, these strategies may lead to colonization by other microorganisms. Therefore, the usage of antibiotic agents, which kill most microorganisms but leave other biofilm components and persister cells behind, must then be addressed[18].

Since biofilms are formed on solid surfaces, in most clinical trials, the test biofilm targeting approaches have been focused on indwelling medical devices. Current strategies can be divided into two groups: Physical-mechanic and surface-coating approaches. The first ones include high-velocity spray and jet irrigators that disrupt and remove the biofilm. The last ones are based on surface impregnation with antibiotics for preventing biofilm formation. However most approaches still use conventional antibiotic-based therapy.[18]

Biofilm infections are not easily treated with existing antimicrobial approaches, because the biofilm recalcitrance is a consequence of its complex physical and biological properties[18]. On the other hand, the number of infections by multidrug-resistant bacteria (MDR) is rising and there is a lack of new antibiotics in development. This means that inhibiting the initial formation of the biofilm, by disrupting cell-to-cell communication, or quorum sensing is the most promising strategy for treating biofilm infections.[32].

Additionally, QS inhibition molecules do not affect the normal growth of the bacteria. Therefore, they do not create any evolutionary pressure for the emergence of MDR bacteria. The QS inhibitors have thus usually a longer functional shelf life than modern antibiotics when it comes to treating diseases caused by pathogenic bacteria[33].

1.2 Quorum Sensing

1.2.1 Definition

Quorum sensing is a process of cell-to-cell communication that relies on the production, detection and response to extracellular signalling molecules called autoinducers. It was discovered in 1979 by Nealson and Hastings, who characterized bacterial communication by studying the bioluminescence mechanisms of *Vibrio fisheri and Vibrio harveyi*[34]. QS allows bacteria to synchronously adjust their gene expression in order to alter their behaviour in response to changes in population density and surrounding bacterial community. It controls processes including bioluminescence, virulence factor production and biofilm formation[35].

QS signalling plays an important role in biofilm formation, in such a way that specific QS signalling blockage is an effective way to prevent biofilm formation of most pathogens[36]. It also reported to regulate various virulence factors which help bacteria to evade the host's immune response and cause pathological damage. Studies have shown that in the case of Gram-negative *Pseudomonas aeruginosa* virulence factors such as pyocyanin, elastase, lectin and exotoxin A are regulated by QS systems[37]. The same applies to Gram-positive *Staphylococcus aureus* virulence factors such as hemolysin, protein A and enterotoxin[38].

Two proteins are required for the QS system in Gram-negative bacteria. One is responsible for the production of the signalling molecule, the autoinducer, and the other responds to the autoinducer[39]. On Gram-positive bacteria, the QS system consists of the signal molecule and a two-component regulatory system that has a membrane-bound histidine kinase sensor and an intracellular response regulator[40]. There are several classes of autoinducers based on molecular features. These include acyl homoserine lactones, autoinducing peptides and autoinducer-2[39].

1.2.2 Autoinducers

1.2.2.1 Acyl Homoserine Lactones

N-acyl homoserine lactones are the main autoinducer used by Gram-negative bacteria, with over 70 species known to communicate via AHL-mediated QS. Specificity is accomplished via variation in the length and oxidation state of the acyl side chain. The native AHL used by bacteria are the L-isomers, whereas the D-isomers are biologically inactive[24]. In *V. fischeri,* the *luxI* gene is activated to produce the AHL synthase enzyme LuxI. Once the AHL concentration in the extracellular environment exceeds a certain threshold, they diffuse across the cell membrane and bind to specific QS transcriptional regulators, the LuxR receptor. Since the AHL can freely diffuse across the cell membrane, the total AHL concentration is proportional to the total bacterial concentration. This enables a population-density control of gene expression[24,41]. Investigation in other Gram-negative bacteria has shown the existence of homologous systems to LuxI and LuxR proteins for AHL synthesis and response[42].

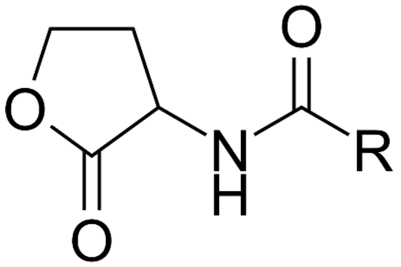


Figure 1 - Conserved chemical structure of acyl homoserine lactones.

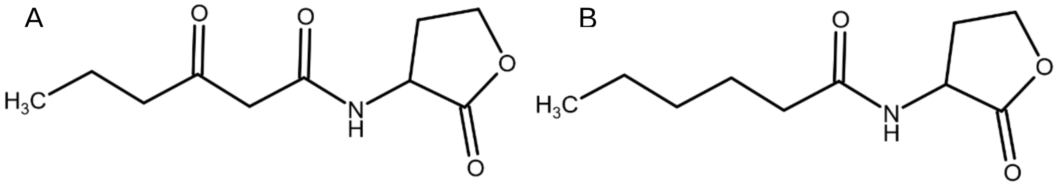


Figure 2 - Chemical structures of: A- 3-oxo-C6-HSL, the AHL for Vibrio fischeri*; B- C6-HSL, a AHL for Chromobacterium violaceum.*

1.2.2.2. Autoinducing Peptides

Autoinducing peptides (AIPs) or quorum sensing peptides (QSP) are post-transcriptionally processed small molecules used by Gram-positive bacteria as autoinducers. Although AIPs are the more common autoinducers in Gran-positive bacteria, they may not be exclusive. In fact, small molecules known as γ-butyrolactones have been identified as signal molecules in some species of *Streptomyces*[40,43].

AIPs are secreted by membrane transporters. When the environmental concentration reaches the threshold, the signal molecules bind to a bicomponent, membrane-bound, histidine kinase sensor. After phosphorylation, this originates a change on the target gene expression[41].

Many AIPs possess hydrophobic domains, which have been postulated to play an important role in the hydrophobic interactions between the ligand and the receptor. Therefore, they are crucial for activity[44].

1.2.2.3. Autoinducer-2

Both Gram-positive and Gram-negative bacteria share this universal quorum sensing mechanism, which involves the production of a small molecule named autoinducer-2 (AI-2). AI-2 molecules are derived from the precursor molecule (S)-4,5,dihydroxy-2,3-pentanedione (DPD), with the DPD synthase enzyme being found in over 55 bacterial species[45].

When the extracellular concentration of AI-2 reaches the threshold, a signal transduction cascade is activated. AI-2 is imported by membrane transporters inside the cell, where it is phosphorylated by kinases and binds to repressors and activators of relevant genes[46].

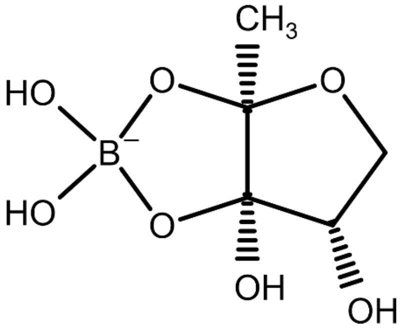


Figure 3 - Chemical structure of Autoinducer-2.

1.2.2.4 Other signalling molecules

There are less common molecules that can also act as signalling molecules in specific bacteria. Cis-2-decenoic acid, was shown to inhibit biofilm development in *P. aeruginosa* and to induce the dispersion of established biofilms formed by various bacteria, including *E. coli* and *S. aureus*[47]. Other fatty acid, cis-11-methyl-2-dodecenoic acid, also known as diffusible signal factor, is able to disaggregate cell flocs formed by *Xanthomonas campestris*[48]*.*

Some D-amino acids are thought to be a native signal for biofilms disassembly in *B. subtilis* and inhibit biofilm formation in fresh cultures of *B. subtilis, S. aureus* and *P.*

aeruginosa. The inhibitory activity of these amino acids is thought to be due to a disruption of the connection between the extracellular matrix protein and the cell[49].

1.2.3 Quorum sensing in different bacteria

1.2.3.1 *Staphylococcus aureus*

S. aureus is a leading case of nosocomial infections worldwide, being the agent of a wide range of diseases. Many of these, including endocarditis and osteomyelitis, appear to be caused by biofilm associated *S. aureus*[50].

The QS system in use is encoded by the accessory gene regulator (agr) operon. The *agrD* gene encodes for the precursor of the *S. aureus* AIP, AgrD, which is processed into the autoinducing peptide by AgrB. AgrA and AgrC comprise a two-component regulatory system that responds to this autoinducer. Signalling via this system, together with other regulatory elements such as SarA, results in elevated intracellular concentrations of a small, non-coding RNA called RNA-III. When the AIP concentration reaches the threshold level it will bind AgrC leading to the expression of RNA-III. This RNA will then down regulate genes, which encode adhesins required for biofilm formation and increase the expression of secreted virulence factors. The RNA-III activating protein (RAP) activates its target, TRAP, via phosphorylation. This induces expression of the *agr* operon and increases cell adhesion and biofilm formation[51,52].

1.2.3.2 *Pseudomonas aeruginosa*

P. aeruginosa is one of the most virulent opportunistic pathogens and is responsible for 10 to 20 % of infections in hospitals[53]. It is a leading cause of various acute infections, including ventilator associated pneumonia. Furthermore, it can also cause chronic lung infections in patients with cystic fibrosis, being associated with increased mortality[54]. It is responsible for 17 % of nosocomial pneumonia, 7 % of urinary tract infection, 8 % of quotidian cause surgical-site infection and 9 % of general infection[55]. These bacteria have a complex genome and a large and variable arsenal of virulence factors. The ability to form biofilms provides them with and enormous advantage to establish infections within susceptible hosts[54].

P. aeruginosa possesses several QS systems with the most important being *las* and *rhl,* both associated with AHL[56]*.* Both systems are homologous to the LuxI and LuxR system[42]. For the *las* system LasI synthesizes the autoinducer, N-(3-oxododecanoyl)-l-homoserine lactone, which activates the LasR transcriptional activator. Likewise, in the *rhl* system RhlI synthesizes the autoinducer, N-butyryl-l-homoserine lactone, which activates RhlR[55].

1.2.3.3 *Chromobacterium violaceum*

Chromobacterium violaceum is a large, motile, Gram-negative bacillus which lives on soil and water in tropical and subtropical regions. Although it is considered non-pathogenic, it can act as an opportunistic pathogen for animals and humans. It enters though broken skin, by contamination with soil or stagnant water[57]. There have been reports of it causing localized skin and soft tissue infection and systemic or invasive infection. These including necrotizing fasciitis, visceral abscesses, osteomyelitis and central nervous system disease[58]. Infections due to *C. violaceum* are very rare with less than 150 published clinical reports, but are associated with high mortality[59].

This bacterium is known for the production of a natural violet pigment with antibiotic properties, known as violacein, whose production is regulated via quorum sensing[60]. Since this QS-regulated trait is an easily observable and quantifiable trait, *C. violaceum* has been widely used as a model organism for QS research[61]. Other phenotypes which are known to be regulated by QS include biofilm formation and the production of chitinase and cyanide[61,62].

The QS system in *C. violaceum* is homologous of the LuxI/LuxR system, with the AHL synthase being CviI and the transcriptional activator being CviR.

The *C. violaceum* virulence is controlled by QS system, as indicated by experiments carried out on the nematode *Caenorhabditis elegans*. It was proved that AHL synthase antagonists, which replace its natural ligand and induce a CviR conformation that prevents DNA binding, protect this nematode from *C. violaceum*-mediated killing[60]. Strain ATCC 31532 produces the autoinducer N-hexanoyl homoserine lactone (C6-HSL) while strain ATCC 12472 produces 3-hydroxyl-C10-HSL[60,63].

Violacein is synthesized by the producers of the *vioABCD* operon, being produced by the fusion of two tryptophan resiudes[64]. Using *vioA* promoter mutations, an ideal DNA binding site for CviR was defined as CTGNCCNNNNGGNCAG. This allowed the discovery of other genes regulated by CviR. These include genes encoding a guanine deaminase, an extracellular chitinase, a protein with a role in type VI secretion, a transcriptional regulator possibly in Multiple Antibiotic Resistance Regulator family and the autoinducer synthase CviI[65].

CviR, like other LuxR-type proteins, functions as a homodimer, with each monomer consisting of a ligand-binding domain and a DNA-binding domain[66].

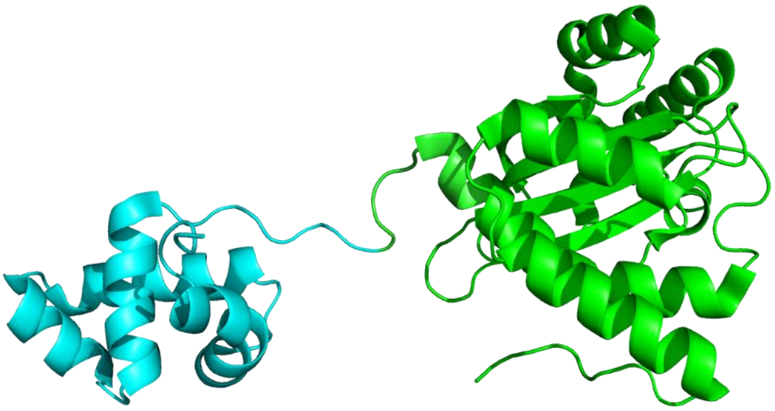


Figure 4 - Structure of CviR represented in New Cartoon. The ligand-binding domain is coloured in green and the DNA-binding domain is coloured in blue.

When an antagonist is used, the dimer adopts a closed, crossed-domain conformation, in which the DBD each monomer is positioned below the LBD of the opposite monomer. This conformation separates the two DNA binding helices in way incompatible with high-affinity DNA binding, as can be seen in figure 4[67].

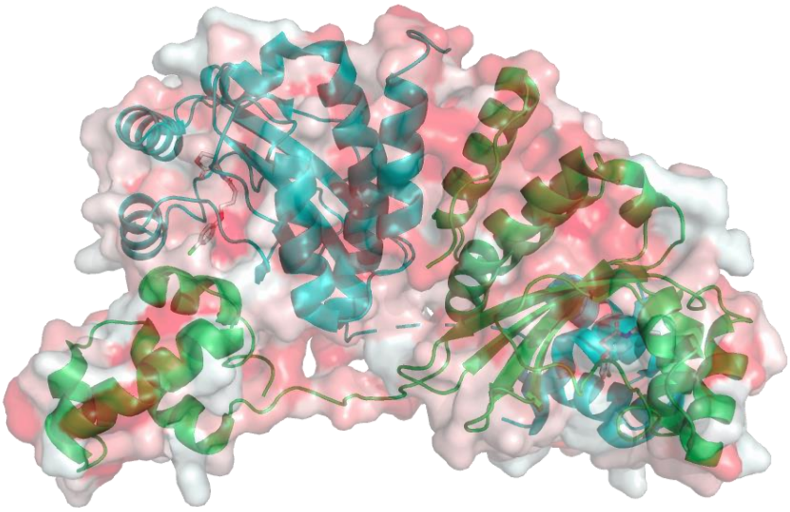


Figure 5 - CviR homodimer after bonding an antagonist. Both chains are represented in the New Cartoon format with chain A is coloured in blue and chain B is coloured in green. The overall protein is represented using the surface representation.

The CviR from both strains is 87% identical in amino acid sequence, with its difference in autoinducer being partially due to a naturally occurring serine instead of a methionine at position 89. This is a key residue that occupies the opening of the ligand-binding pocket. The importance of Met 89 is reinforced by the fact that, when using antagonists, the side chain swings away. The extension of this movement increases with strength of the antagonist[67].

Although it is not often involved in human infections, it is regarded as an emerging pathogen. On the other hand, since *C. violaceum* is used as a model QS system, it is very useful for the discovery of drugs able to inhibit infections[61]. Therefore, the research of novel molecules able to supress the QS system in this bacterium could be very useful, not only for the treatment of *C. violaceum*-mediated infection, but also for the discovery of molecules that inhibit other LuxR-type receptors by analogous mechanisms.

1.3 Databases for the Study of Biofilms

Databases for the Study of Biofilms: Current Status and Potential Applications
Fábio G. Martins, André Melo and Sérgio F. Sousa
Under submission

1.3.1 Introduction

A database is an organized collection of related information that can be stored and accessed electronically. The significant growth of some widely used databases such as the Protein Data Bank PDB[68], ChEMBL[69] or PubCHEM[70], demonstrates how important they are for current research. The number of available databases has also rapidly increased over the years. In 2005, the Nucleic Acids Research Molecular Biology Database Collection reported a total of 719 databases, while in 2020 the number of databases is over twice as large, with a total of 1637 databases[71,72].

Considering the prominence of biofilm research and the importance of databases in this information age, the development of specialized databases designed to handle information related to microbial biofilms was a logical step. Currently there are 7 online databases (Table 1) on the subject of microbial biofilms: (1) “Quorumpeps” was created in 2012 and provides an overview of reported QS signalling peptides and their derivatives[73]; (2)“BiofOmics” was created in 2012 and was the first public web platform for the systematic and large-scale compilation, processing and analysis of biofilm data from high-throughput experiments[74]; (3) Biofilm-active AMPs, or “BaAMPs”, was created in 2015 and it was the first database dedicated to antimicrobial peptides (AMPs) with antibiofilm activity, with the goal of building an open access resource that gathered all experimental information about antibiofilm AMPs[75]; (4) QSPPred was created in 2015 as a platform for predicting and designing quorum sensing peptides[76]; (5) design Peptides Against Bacterial Biofilms, or dPABBs, was created in 2016 and functioning also as web server for predicting antibiofilm peptides and creating mutants with improved activity[77]; (6) "aBiofilm" was created in 2017 gathering chemical, biological and structural data for various anti-biofilm agents[78]; (7) The Biofilms Structural Database, or BSD, was created in 2020 as an open-access collection of all known structures of proteins involved in biofilm formation. It has the goal to aid the research of novel antibiofilm drugs and the understanding of the structure and activity of proteins participating in biofilm formation[79].

Table 1 - Overview of all currently available biofilm databases.

Name	Year	Reference	Website	
Quorumpeps	2012	73	http://quorumpeps.ugent.be	Information on Quorum sensing peptides
BiofOmics	2012	74	http://www.biofomics.org/	Standardized biofilm experimental data
BaAMPs	2015	75	http://www.baamps.it/	Antibiofilm antimicrobial peptides experimental data
QSPPred	2015	76	http://crdd.osdd.net/servers/qsppred	Quorum sensing peptide prediction
dPABBs	2016	77	http://ab-openlab.csir.res.in/abp/antibiofilm/	Antibiofilm peptides prediction
aBiofilm	2017	78	https://bioinfo.imtech.res.in/manojk/abiofilm/	Various data about several anti-biofilm agents
BSD	2020	79	https://biosim.pt/biofilms/	Structural information about proteins involved in biofilm formation

1.3.2 The Quorumpeps Database

Quorum sensing plays a major role in biofilm formation and is also involved in the regulation of multiple virulence factors. Therefore, the research of QS signalling molecules has received an increased interest. In gram-positive bacteria the quorum sensing phenomenon is driven by the involvement of signalling molecules that are named Autoinducing peptides or Quorum sensing peptides (QSPs)[9,35,73].

With this in mind, Quorumpeps was developed by the Drug Quality and Registration group in Ghent, Belgium. The goal of this database is to document the structures, microbial origin, and functionality of QS-derived signalling peptides[73].

The creators performed an intensive research on several search engines and the relevant information was included in the database. Users can submit further entries through the data submission page. In this page users must enter their name, e-mail, and information about the sequence. This includes the sequence, trivial name, SMILES, molecular formula, the origin of the peptide, information about its functionality, citation, and any further comments the user desires to submit[73].

Each entry on the database features the one-letter amino acid sequence, the Quorumpeps molecule ID and the molecular formula. Furthermore, there are five further tabs of information. The first one, Chemical information, features the above-mentioned information. It also contains the trivial name, the SMILES code, the molecular weight, the LogP value and the isoelectric point of the peptide. The species origin, as the name implies, contains the origin of the peptide. For example, originated from a bacterial species, phage-produced, or produced by modification on another peptide. The functionality tab features the method, or methods, which were used to study their QS activity. The links tab includes links to other peptides available in the database which are related to the selected one. These are divided in peptides that are active and synthesized by the same species or active on the same receptor. The peptides which are active on the same receptor are further divided into agonists or antagonists. Finally, the literature section features a Quorumpeps publication ID. Each publication ID features the title of the article, the authors, the year of release, the journal on which it was published, links to all other peptides available on this database that are related to this article, and a link to PubMed where the user can download this publication[73].

To find the desired information, the user can search this database by sequence, trivial name, SMILES, molecular formula, receptor, method, and origin of literature. There is also a list of known receptors, methods, and species in order to ease the search. To perform a more detailed search within the obtained results, the user can also use a new keyword on the search box and click “refine”. If the user desires to directly compare multiple entries, this can be done by selecting the desired peptides and clinking “compare results”[73].

This database can be useful for multiple experiments. The user can find the QSPs for a variety of different targets. This can be used in a laboratory setting, where this information allows the researcher to find specific QSPs for their desired receptor. On an in-silico experiment, one can use this database for conservation analyses via multiple sequence alignment. The SMILES available for each entry can easily be converted into a 3D structure, which can then be used, for example, on molecular docking experiments.

Quorumpeps can also be used for the development of other databases or even as a positive or negative dataset in the development of machine learning based predictors. In fact, this database was used as one of the sources for the development of other databases also discussed in this article. In fact, it was used during the development of dPABBs, QSPpred and aBiofilm[76–78].

In their work about quorum sensing in thermophiles, Kaur et al.[80] use Quorumpeps to find which thermophiles have been reported to exhibit quorum sensing though the use of quorum sensing peptides.

Rajput and Kumar, in their 2017 article about LuxI and LuxR homologs in Gram-positive bacteria[81], used 51 peptides obtained from the Quorumpeps database. They then further analysed and predicted their activity using QSPpred.

In general, Quorumpeps presents a structured overview of all reported quorum sensing peptides and their derivatives. It provides an easy procedure to use search engine, and information on a large variety of protein targets belonging to different species. Currently, it lacks the ability to download the structure of each peptide, which could make the study of multiple peptides an easier process. Overall, this database can be very useful in multiple settings, and with the updates from the authors and user made submissions, the database remains a useful platform for research related to QSPs.

1.3.3 The BiofOmics Database

The emergence of high-throughput technologies led to biofilm studies becoming more and more data-intensive. Controlling and managing this information became essential, which led to the development of a computational tool to manage this information: BiofOmics. This platform was developed in a joint effort between the Institute for Biotechnology and Bioengineering at the University of Minho and the Laboratory for Process, Environmental and Energy Engineering of the Faculty of Engineering at the University of Porto. Since most biofilm information remained with its original researchers and often involved different data processing, the goal of BiofOmics was to promote data interchange across laboratories. BiofOmics achieves this by collecting, storing, standardizing and providing open access to biofilm research data[74].

Research groups across the world are encouraged to submit their data into BiofOmics in order to populate the database. This is done via an online data submission interface. This allows the researchers to fully characterize the experiment’s goals, operational environments, and results. Considering that one of the goals of this platform is to provide access to standardized biofilm data, BiofOmics provides a three-step protocol to describe the experiment, standardize the data and upload it to the database. Data is standardized via Microsoft Excel worksheets, which researchers use to “translate” the terminology to what is used in this database. Finally, curation tools check the data for any typos, non-compliant data, structuring and data inconsistencies. This originates a validation report which is sent to the researcher for any corrections[74].

BiofOmics allows its users to search the entire database for desired information by indicating the biofilm-forming microorganism, biofilm-forming device, growth medium, adhesion surface or desired antimicrobial product. This search originates a list containing a number of potentially relevant studies from which the data and associated publications of any study can be accessed[74].

This sharing of standardized data enables an easier distribution of information and allows the comparison of multiple experiments. This can ease research in this field in multiple ways. These include allowing the search for similar experiments, the raise in awareness for relevant but under-reported areas, and the statistical analysis of experimental robustness and reproducibility.

The BiofOmics database has been mentioned in multiple articles, being acknowledged for their effort in offering the first public systematic and standardized collection of guidelines, experiments, and biofilm data[82–84]. By helping the interchange of biofilm data, this platform can promote the collaboration across laboratories. This helps researchers in searching for similar experiments and developing standardized protocols for biofilm-related studies[74].

1.3.4 The Biofilm-active AMPs Database

Antimicrobial peptides (AMPs) are a large and diverse set of molecules which are part of the innate immune system. These molecules show potential for the development of new therapeutic strategies to prevent biofilm formation or to destroy existing structures[75,85]. While many databases had been created for collecting AMP sequences, none was focused on AMPs active for microbial biofilms. The BaAMPs (Biofilm-active AMPs) database was created to collect data on AMPs tested on microbial biofilms. This database was developed at the Istituto Nanoscienze-CNR and Scuola Normale

Superiore, the Center for Nanotechnology, Istituto Italiano di Tecnologia, and the Dipartimento di Ricerca Traslazionale e delle Nuove Tecnologie in Medicina e Chirurgia, Università di Pisa[75].

The BaAMPs database contains experimental data extracted from 86 articles, including 162 AMP sequences and 422 experiments. All information was obtained by a detailed literature search and all sequences were compared with the corresponding ones in public databases in order to prevent errors. This database includes a peptide list and an experiment list[75].

The peptide list arranges the information of each entry into two sections, the general characteristics and the experimental information. The general characteristics and attributes section contains the name of the AMP, sequence, source, and several characteristics such as the isoelectric point, molecular weight and hydrophobicity. The experimental section contains a list of experiments categorized by target organism, featuring a link to the corresponding experiment page. The experimental list contains information about the procedure used and the results. This information includes the target microorganism, the method of AMP administration, the stage of biofilm formation evaluated, method used to evaluate the activity, biofilm reduction and the concentration tested. Each experiment page is also linked to the corresponding peptide page. All peptide and experimental pages additionally contain references to the original article[75]d.

Besides the database, BaAMPs contains a peptide properties calculator which calculates and compares crucial physicochemical properties necessary for anti-biofilm activity. Users can also submit simple sequences to the NCBI BLASTP server to search for similar sequences[75].

The database can be expanded by user contribution. After registration, any user can submit new sequences and experimental data. The database administrators will validate any new entry before it becomes available in the database, to guarantee the desired levels of accuracy[75].

With all the different peptides available in BaAMPs, this database can be used to find peptides against a specific target. They can then be further tested or used as a benchmark for comparing novel molecules. This database can also be useful for conservation analysis, or, using the BLAST section of this website, to find similar peptides to the selected one.

The availability of experimental information about the activity of multiple peptides against multiple targets makes BaAMPs a very useful tool for the development of machine-learning tools to predict antibiofilm AMPs. dPABBs, a predictor further explored in a later section of this paper, used peptides obtained from BaAMPs as the positive dataset for training its machine learning model. Another predictor developed with the help of BaAMPs was the one developed by Gupta et al.[86]. The developers used the experimentally validated anti-biofilm AMPs available on the BaAMPs database to develop machine learning based prediction models (Support-Vector Machine) for new biofilm inhibiting peptides. 178 unique AMPs were used as a positive dataset and compared with randomly generated peptides. Using a frequency analysis, it was clear that positive charges and aromatic amino acids were more common on anti-biofilm peptides. All the models built displayed over 90 % accuracy in the identification of biofilm inhibiting peptides.

In 2019, Vergis and colleagues[87] evaluated the effectiveness of indolicidin against multi-drug resistant enteroaggregative Escherichia coli strains in the Galleria mellonella larval model. Indolicidin, which was retrieved from BaAMPs and synthesized commercially, was shown to fully eliminate the bacteria, while being safe to the eukaryotic cells.

In their study concerning the usage of antimicrobial peptides to provide resin composite restorations with a 2-tier protective system, Moussa, Fok and Aparicio[88] used BaAMPs to obtain the physical and chemical properties of the peptide used in this study, the GL13K peptide. The usage of this peptide gave antimicrobial properties to the coating which is expected to increase the durability of resin composite restorations.

In this present version, BaAMPs does not provide the possibility of batch downloading for multiple peptides. Although the user can search for an experiment by its target organism, currently it is not possible to search peptides by their target organism.

Nevertheless, BaAMPs provides a useful toolbox for researchers and can help in the study and design of novel AMPs with anti-biofilm activity.

1.3.5 The QSPpred predictor

While Quorumpeps has information on multiple QSPs, it does not have any predictive ability. To answer this need, QSPpred was developed in the Bioinformatics Centre of the Institute of Microbial Technology, Council of Scientific and Industrial Research in India[76].

The authors developed Support-Vector Machine models for the prediction of the peptides. For the positive dataset, the authors of this work used QSP obtained from Quorumpeps and PubMed. Otherwise the negative dataset was obtained from UniProt[76].

The QSPpred website hosts these SVM based prediction models, named QSPepPred, QSPepDesign and QSPepMap. QSPepPred analyses one or multiple input peptides and quickly predicts if these are quorum sensing peptides. The results show the score for each peptide and if it is classified as a QSP or Non-QSP. The results are colour-coded, with purple indicating a more effective, while green indicates a less effective peptide. The user can also access a graphical representation of multiple physicochemical properties of the peptide such as amino acid composition and hydrophobicity. QSPepDesign allows users to design quorum sensing peptides by developing all possible single point mutants of a given peptide sequence. After generating the mutants, it then predicts their activity, offering similar results to those obtained in QSPepPred. QSPepMap was designed to identify potential regions of a protein which could have quorum sensing activity. Finally, this platform also features various analysis tools: QSMotifScan, that allows users to scan possible QS motifs in a sequence; MutGen that allows users to create custom mutations in any sequence; PhysicoProp, that provides physicochemical information on any sequence; ProtFrag, that creates multiple fragments from one input protein. All the peptides used for the development of the predictive models can be downloaded from the dataset section of this website. All obtained results can be downloaded in excel format[76].

All the tools available in this platform can be fitted into customized workflows. For exemple, QSPepMap can be used for finding new QSPs from the proteome of any bacteria. Alternatively, a peptide sequence can be fragmented by ProtFrag and the resulting peptides can be predicted using QSPepPred. In an experiment in which it is necessary to have multiple QSPs, one can use QSPepDesign to find new QSPs from an initial peptide. Before advancing to an experimental setting, QSPepPred can be used to predict the activity of library of multiple peptides, reducing the number of peptides to be tested.

QSPpred was used on the development of the dPABBs predictor to obtain a series of QSPs that were used as the negative training set for this predictor[76].

As mentioned previously Rajput and Kumar, in their article about LuxI and LuxR homologs in Gram-positive bacteria[81], analysed and predicted the activity of multiple QSPs obtained from Quorumpeps using QSPpred.

In 2016, Pang and colleagues published an article reporting the identification of quorum sensing signal molecule of *Lactobacillus delbrueckii subsp. Bulgaricus*[89]. Candidate molecules were analyzed using QSPpred, to predict their activity and obtain their physicochemical properties.

Overall, QSPPred is easy to use and contains multiple tools that can be used in the identification of new quorum sensing peptides used in biofilm formation or virulence mechanisms. It also can help with the discovery of new QSPs which can work as therapeutic targets. Thus, QSPpred provides a helpful platform for this area of research.

1.3.6 The design Peptides Against Bacterial Biofilms Predictor

The dPABBs (design Peptides Against Bacterial Biofilms) was developed at the Open Source Drug Discovery (OSDD) Unit, Council of Scientific and Industrial Research, in New Delhi, India. As was the case with BaAMPs, this platform is focused on antimicrobial peptides with antibiofilm activity. Whereas BaAMPs is mostly concentrated on the documentation of experimental data, dPABBs has the goal of predicting if a peptide is active against biofilms and to display several mutations that can improve the antibiofilm activity[77].

The developers of this platform built six SVM and Weka-based models using machine learning tools. The positive dataset used to train these models consisted of 80 AMPs obtained from the BaAMPs database. As for the negative dataset, the developers considered that quorum sensing peptides, which exist and function within the biofilm, would have a contrasting set of properties to peptides that would disrupt biofilms. Therefore, the negative dataset consisted of 88 QSPs obtained from QSPpred and cross referenced with Quorumpeps[77]. All molecules used for the datasets can be downloaded in the Download section of the website.

This platform consists of four different modules: Peptide, Protein, Batch and MultiModel. In the peptide module, the user enters the desired peptide sequence, selects the preferred model (SVM or Weka), selects the appropriate physicochemical properties to be displayed and then can run the analysis. This analysis will assess the antibiofilm activity of a single peptide sequence and generate mutants with successive amino acids substitutions. The user can then select those mutants with higher SVM scores or Weka probabilities and/or better physicochemical properties. In the protein module the user can input a protein sequence to analyse all the possible overlapping peptide fragments and therefore identify those presenting anti-biofilm activity. Before running the analysis, the user can select the desired length of the peptide fragment, which model is going to be used and what physicochemical properties will be displayed. Using the batch module, the user can screen antibiofilm peptides from a peptide library in FASTA format in a single run. This procedure enables to obtain similar results, as in the peptide module, for each of the screened sequences. Finally, the MultiModel module allows the usage of multiple models at the same time. This enables the user to identify which peptides have been predicted to be active against biofilms by many different models. dPAABs also features a list of FDA Approved peptides and the respective prediction made by each different model[77].

This application is useful for predicting if a previously untested peptide or family of peptides is active against biofilms, to find potential antibiofilm peptides from the proteome of a specific organism or to generate multiple mutants starting from a single peptide. Similar to other predictors, it can also be used to predict the activity of multiple peptides before advancing to the laboratory, minimizing costs.

In their study about the effect of amino acids substitutions on the biological activity of certain AMPs, Chegini et al.[90] used the dPABBs database to evaluate the selected template AMP, Magainin II, and generated mutations that would improve the antibiofilm activity.

Marimuthu et al.[91] analysed temporin AMPs and their interactions with the Middle East Respiratory Syndrome-Coronavirus. One step of this work involved predicting which of the temporin AMPs had antibiofilm activity. This was accomplished using dPABBs, which predicted high antibiofilm activity for some temporins.

In 2017, Leoni and colleagues[92] performed an in-silico study of the potential biological activity of myticalins, a novel family of AMPs. The dPAABs database predicted a negligible anti-biofilm potential for most peptides of this family.

The dPABBs allows its users to not only predict the antibiofilm activity of the desired sequences, but also offers mutations that can have more desirable properties. On the other hand, presently there is no way to easily download the results, something that can be done on other predictors, such as QSPpred. Nevertheless, this database is a useful platform in order to predict, identify and optimize antibiofilm peptides.

1.3.7 The aBiofilm Database

The large number of studies involving the experimental testing of antibiofilm agents reported in the literature over the years, prompted the need to develop a specialized platform gathering all the information on the molecules in an easily accessible way. To address this need, “aBiofilm” was developed in the Bioinformatics Centre of the Institute of Microbial Technology, Council of Scientific and Industrial Research in India. This platform contains a database, a predictor, and a data visualization module[78].

The aBiofilm database contains biological, chemical, and structural information on 5057 entries belonging to 1720 unique antibiofilm agents, targeting 140 microorganisms. All information was obtained by an intensive literature search which resulted in the before mentioned 5027 entries, which were obtained from 526 articles and manually curated. The biological information reported in the aBiofilm database includes the type of anti-biofilm agents, organism and strain targeted, concentration of agent, percentage of inhibition, stage of biofilm targeted and the mechanism of action. The chemical information includes the IUPAC name, SMILES, molecular formula, molecular weight, InChI and Lipinski’s rule of five. This information was extracted from various chemical repositories. The structural information contains the 2D and 3D representation of every molecule in the database[78].

This database is organized into six different sub-categories. These sub-categories are anti-biofilm agents, type of anti-biofilm agents, target organism, type of target organism, preliminary assays, and the Journal name. The anti-biofilm agents sub-category, presents a list of all available molecules. Selecting one of the agents leads to a different page where all entries in this database featuring this compound can be seen. The AntiBiofilm_ID of a specific entry directs to a page featuring all biological, chemical, and structural information about this entry. The other sub-categories are organized in a similar way, the only difference being the initial list being associated to each sub-category. The user can also search the database in a specific search page. The search can be done by Antibiofilm_ID, antibiofilm agent, SMILES, Antibiofilm agent type, Organism, Preliminary assay and/or PubMed ID[78].

aBiofilm also contains a Quantitative Structure–Activity Relationship (QSAR) based predictor for the inhibition efficiency of any chemical against biofilm. The Support-Vector Machine model development was based on a curated set of 492 chemicals. These chemicals were divided into 450 molecules for training and 42 for independent validation. Any user can input a chemical in SMILES, sdf or mol format, or draw the chemical in a

JSME window. The result is a table which includes the SMILES code of the input molecule, the predicted anti-biofilm efficacy, 2D and 3D structures and some general properties. The molecule is also searched in the database for any similar compounds[78].

The data visualization module is divided into three sections: (1) A CIRCOS plot displaying the organisms and the step of biofilm formation targeted by the anti-biofilm compounds present in the database; (2) cytoscape based interaction maps highlighting the relationship between the anti-biofilm agents, inhibition efficiency and stage of biofilm formation for each targeted organism; (3) a chemical clustering was used to represent the diversity of the compounds available in the database[78]. There are several studies that have taken advantage of the aBiofilm database capabilities.

The large number of compounds available in this database can have multiple uses. These molecules can be used as a training set for the development of new machine-learning tools to predict new antibiofilm agents. It can also be a very good source for the development of new databases regarding antibiofilm or antimicrobial agents. This database can also be used on molecular docking and virtual screening protocols. The compounds available in this database can be used either on the optimization of the protocols, being used as sets of active and/or inactive molecules for specific organisms, or on the virtual screening process itself. As with other predictors, the aBiofilm predictor can also be used to predict the activity of several agents before advancing to further experiments. This early prediction can be done before further computational studies or before advancing to screening procedures, reducing the number of compounds to be evaluated.

Tiwari[93] studied the secondary metabolites produced by nosocomial pathogens, to discover if these molecules contributed to their survival over other bacteria in the hospital setup. In one of the steps of this study, the aBiofilm predictor was used to predict the anti-biofilm activity of 23 antimicrobial secondary metabolites, with multiple molecules exhibiting a high or very high predicted anti-biofilm activity. This study concluded that these nosocomial pathogens carry antimicrobial secondary metabolites, and that most of them have anti-biofilm activity.

Almeida et al.[94] performed virtual screening studies involving the quorum sensing receptor present in Salmonella. Several plant compounds and nonsteroidal anti-inflammatory drugs were docked into SdiA (Suppressor of division inhibitor receptor). SdiA is a homologue of LuxR, a transcriptional regulator from Vibrio fischeri, involved for the quorum sensing mechanism[1]. Besides the molecular docking studies, the aBiofilm predictor was used as an indicator of the predicted anti-biofilm activity of each molecule. Most molecules tested in this study bind to at least one of the three structures of SdiA available. Many of these molecules were shown to have higher binding affinity than the native inducers of the quorum sensing mechanism and were predicted to have high anti-biofilm activity.

In addition, during the development of MDAD, a comprehensive microbe and drug association database, their creators used the aBiofilm database as one of the sources for the entries available in this database[95].

In general, the aBiofilm database is simple to use and features a large number of compounds, targeting multiple organisms. Even though there is currently no way to download compounds, either one by one or in batch, through its database and predictor, aBiofilm provides a helpful platform for researchers working in the development of anti-biofilm agents.

1.3.8 Biofilms Structural Database

Over the last decade, there has been a very significant increase in the available structural information on proteins and enzymes involved in biofilm formation for many bacteria. Hundreds of crystallographic structures on potential targets for the development of new anti-biofilm drugs became available on the Protein Data Bank. This shift from the cellular to the molecular field was also accompanied by the availability of a large amount of chemical and molecular data on new active molecules on databases such as the ChEMBL and BindingDB. This, together with other biological or bioinformatics-oriented databases offers new possibilities in the application of techniques such as virtual screening[96–98], protein-ligand docking[99–101], QSAR models[102,103], and molecular dynamics[104,105], for the research of new anti-biofilm agents. In order to aid this research, the Biofilms Structural Database[79] was developed in 2020 at the University of Porto. BSD contains all available structural information on proteins involved in biofilm formation. This database can aid researchers both in the study of proteins involved in biofilm formation and in the study of substrate recognition, with the ultimate goal of helping the development of new anti-biofilm agents.

BSD contains currently a total 425 PDB entries, which correspond to 133 unique proteins and 93 ligand molecules, from 42 bacteria. All information available on the database was obtained through literature search and all information and structures was manually checked and validated[79].

Each entry contains the respective PDB code, three-dimensional representations for the representative fragments of the biomolecular (full protein, ligand and surrounding amino acid residues), an interactions map of the ligand, and general information about the structure. The general information includes the category of the protein, the mechanism of biofilm associated with it, the bacteria and specific strain, the ligand, the method by which the structure was obtained, resolution, year of deposition on the PDB, and the DOI of the corresponding publication. When available, each entry contains links to the corresponding sections in the ChEMBL, BindingDB, ExPASy, KEGG, and UniProt databases. This way, more information on each entry is readily available. Clicking on each ligand presents information including molecular weight, SMILES, molecular formula, InChI, PDB code, among other properties. Any selection can be downloaded as a comma-separated values (CSV) file, or as a collection of PDB structures. New entries can be suggested on a dedicated submission page. In this page, the user must enter their e-mail and the PDB code of the suggested structure. The user can also add further information. This includes the protein associated with this structure, the mechanism in which it is involved, the category, autoinducer type, bacteria, strain, gram-type, and DOI, as well as ChEMBL, BindingDB, ExPASy, KeGG and UniProt links[79].

This website also contains a page with statistical analysis of several indicators. These include bacteria, gram-type, type of autoinducer, resolution of the structure and the number of deposited structures per year[79].

BSD can be useful for a variety of computational techniques, particularly those requiring a molecular representation or description of the proteins involved. The user can easily download all available PDB structures for a specific protein, and use them for molecular docking, virtual screening, and molecular dynamics procedures. By providing links to ChEBML and BindingDB, the user can find multiple active compounds against the desired target. This can be useful for the development of QSAR or for optimizing virtual screening protocols.

BSD helps researchers to visualize, explore and understand biofilm targets. Therefore, it is a useful tool in the development of effective antibiofilm agents. While still very recent, this database is expected to play an important role in this field, by connecting molecular drug discovery and biofilm research.

1.3.9 Future Challenges

While these seven databases are certainly useful tools for biofilm research, at different levels and with different purposes, much remains to be done. There is a marked challenge for combining decades of research at the fundamental level, with the recent advances in genomic, proteomics, systems biology and high-throughput techniques, while preparing the future, in which big data and machine-learning will play an ever-increasing role. This requires much larger and more comprehensive databases, in a dimension and diversity several orders of magnitude higher than the present available alternatives.

The design and development of biofilm-related databases constitutes a specially demanding multidisciplinary challenge. This involves researchers from very diverse fields, such as: environmental sciences[106,107], chemistry[108,109], applied biology[110,111], evolution[112,113], ecology[114,115], molecular biology[116,117], medicine[118,119] and dentistry[120,121]. Biofilm data is dispersed through a number of data sources, targeted towards specific audiences, built with a different aim, and anchored on different premises. Combining this information in a structured way, through conventional databases, requires the use of carefully designed and specifically oriented architectures. This has limited the development of such type of databases. The recent development of unstructured big data databases will certainly be an interesting alternative for many of these problems.

The availability of open access public databases, as well as the expected development of newer and larger ones, is supposed to play an important role for the success of biofilm related research. These databases allow the researchers access to a large body of data, which can be directly applied for studying of individual biofilm targets or identifying new molecules with potential biofilm inhibiting activity. Additionally, they can be used to develop and validate new specifically designed tools and algorithms. Furthermore, they can also promote the exchange of information between researchers of related fields, stimulate cooperation and accelerate the creation of knowledge. These goals can be obtained by enabling incremental research built from previous results and decreasing the number of repeated or irrelevant experiments.

1.4 Protein-Ligand Interactions

In order to perform molecular docking studies, it is necessary to understand the binding model and physicochemical mechanisms involved in protein-ligand binding.

Many proteins function through the reversible binding of other molecules, called ligands. Ligands can be any kind of molecule, ranging from small molecules to a full protein. A ligand binds to the protein at the binding site. This site should be complementary to the ligand in shape, size, charge and hydrophobicity or hydrophilicity. This means that the interaction is specific, with the protein being able to distinguish between thousands of different molecules and selectively interact with only one or a few types[122].

Proteins are flexible entities whose molecular motions cover a wide range of timescales/amplitudes. Protein motions occur on time-scales ranging from 10^{-14} to 10^{-12} s and covering amplitudes ranging from 0.01 Å to more than 100 Å. Several types of motions can be distinguished according to the respective timescales. These include bond stretching, angle bending, constraint dihedral torsions, unhindered surface side chain motions, loop motions, helix coil transitions, collective motions, and protein folding. Specific conformational changes are often indispensable to a protein's function. This is often observed when it comes to the binding of a ligand to a protein, which is coupled by a conformational change in the protein in order to make the binding site more complementary to the ligand[122].

There are three models which have been proposed to explain the protein-ligand binding mechanisms, the lock-and-key model, induced fit and conformational selection. In the lock-and-key model, both the protein and the ligand are rigid, and their binding interfaces are perfectly matched, meaning that only the correctly sized ligand can be inserted into the binding pocket. This leads to a problem, because this model cannot explain when a protein binds a ligand even though their initial shapes do not match. This is explained by the induced fit model. This model assumes the binding site is flexible and interacting with a ligand leads to a conformational change at the binding site. Both the lock-and-key and induced fit model treat the protein as a single stable conformation. However, most proteins are inherently flexible, and the conformational selection model takes this into account. This model proposes that the native state of the protein does not exist as a single, rigid conformation, but as a vast ensemble of conformational states that exist in equilibrium. The ligand is able to bind selectively the most suitable conformational state, shifting the equilibrium towards it. Since all three models have been observed experimentally, all three may exist in a simultaneous or sequential manner, covering a vast spectrum of binding events[123].

In general, the reversible binding of a protein to a ligand can be described as:

$$P + L \leftrightharpoons PL$$

Equation 1- Equation for protein-ligand binding.

This equilibrium is characterized by the association constant (K_a) calculated as:

$$K_a = \frac{[PL]}{[P][L]}$$

Equation 2 - Association constant.

The equilibrium associated with the inverse reaction (dissociation of the protein-ligand complex) is characterised by the dissociation constant (K_d)[122]:

$$Kd = \frac{[P][L]}{[PL]}$$

Equation 3 - Dissociation constant.

Neglecting entropy as a discriminative effect for binding, these equilibrium constants (K_a and K_d) also provides an indirect measure of affinity of the ligand to the protein. For this purpose, the dissociation constant is more used than the association constant. A lower value of K_d corresponds to a higher affinity[122].

The dissociation constant is correlated with the standard molar Gibbs energy of association ($\Delta_a G^0$) by the equation 4, where R is the ideal gas constant (8.3134 J $mol^{-1} \cdot K^{-1}$) and T is the temperature expressed in Kelvin. In standard conditions, the binding of a ligand to a protein is favourable from a thermodynamic point of view when ($\Delta_a G^0$) is negative[122].

$$\Delta_a G^o = -RTlnK_d$$

Equation 4 - Relationship between standard molar Gibbs energy of association and the dissociation constant.

Gibbs energy of association can also be expressed in function of its enthalpic component, $\Delta_a H^0$, and the change in entropic components ($\Delta_a S^0$), as shown in equation 5[122].

$$\Delta_a G^o = \Delta_a H^o - T\Delta_a S^o$$

Equation 5 - Standard molar Gibbs energy of association as a function of its enthalpic and entropic components.

The standard molar enthalpy of association ($\Delta_a H^o$) mainly reflects the energy change of the system in response to the binding of the ligand to a protein. For reversible ligands, this thermodynamic quantity is strongly dependent on the energetic balance between the broken and formed non-covalent interactions. This included van der Waals, electrostatic and hydrogen bonds contributions. These interactions are broken in the protein-solvent and ligand-solvent species. All these individual components can have a positive or negative contribution, and the net standard molar enthalpy change is a result of a sum of all these components[123].

Entropy can be correlated with the information available about the system[122]. It can measure how evenly the heat will be distributed over the thermodynamic system. In an isolated system, the second law of thermodynamics explains that the heat flows spontaneously from regions of higher temperature to regions of lower temperature, reducing the information available in the initial system. The standard molar entropy of association $\Delta_a S^o$ is a global thermodynamic property of a system, which can be calculated as the sum of three compontents[123].

$$\Delta_a S^o = \Delta_a S^o_{solv} + \Delta_a S^o_{conf} + \Delta_a S^o_{r/t}$$

Equation 6 - Standard molar enthalpy of association $\Delta_a S^o$ decomposition.

$\Delta_a S^0_{solv}$ is the component associated with the desolvation/solvation events and usually favours the association process. $\Delta_a S^o_{conf}$ is the component associated with the conformation changes occurring in both protein and ligand upon binding. These components tend to disfavour the association process. $\Delta_a S^o_{r/t}$ is the component associated with the transformation of three translation and three rotational higher-entropic degrees of freedom of the separated species (protein and ligand) in six lower-entropic vibrational degrees of freedom in the protein-ligand complex. This component always contributes unfavourably to the association process[123].

For the binding to occur, the association process must overcome the inescapable entropic penalties such as the negative $\Delta_a S^o_{r/t}$. This can be achieved through large solvent entropy gain or favourable protein-ligand interactions[123].

The Gibbs energy of association is the driving force of the protein-ligand binding process. Therefore, this thermodynamic quantity is used, together with its components, in computer-aided drug design to predict the binding and affinity of a ligand to a specific protein.

2. Computational methods

2.1 Computer-Aided Drug Design

The main purpose of the pharmaceutical industry is to introduce in the market new molecular entities (NMEs). A NME is a medicine containing an active ingredient which has not been previously approved in any form. The process of drug discovery and development is very expensive when it comes to money, manpower, and time. Introducing a NME to the market takes on average 10 to 15 years and it can cost between 800 million and 1.8 billion US dollars[124]. Advances in chemical synthesis allowed the increase of compound databases and the development of high-throughput screening (HTS). HTS is a brute force approach, screening a high number of molecules to find the ones with promising activity. Is has the advantage of not requiring prior knowledge, minimal compound design and often resulting in hit compounds[125]. However, the hit rate is usually low and the lack of understanding of the molecular mechanism can hamper the search of promising candidates[125,126]. Furthermore, the number of NMEs launched into the market has decreased over the years. With this in mind, computer-aided drug design (CADD) has become essential for the preliminary stage of drug discovery, in a process that is more cost-efficient, and it minimizes failures in the final stage[126].

The interest in CADD started to rise in 1981, after an article published in Fortune magazine about drug design using this technique at Merck[127]. However it was only during the last decade that this concept re-emerged as a way to significantly reduce the number of molecules needed to screen for obtaining the same number of lead compounds discovered[125]. This is achieved because compounds that are predicted to be inactive are skipped, and those predicted to be active are prioritized. This way, the cost and workload are reduced, while the lead discovery rate is maintained, when comparing to a full HTS[125].

This field of research has been through a raise of popularity and expansion due to the advances in computational power and software, together with the increased amount of 3D structures of potential drug targets being deposited on the protein databank[128].

CADD is defined by IUPAC as “all computer assisted techniques used to discover, design and optimize compounds with desired structure and properties” and has three main uses: filter large compound libraries into smaller sets of predicted active compounds that can then be tested experimentally; guiding the optimization of lead compounds by increasing its affinity to the target, or to optimize pharmacodynamic and pharmacokinetic properties; building new drugs, one functional group at a time or by joining fragments together[125].

This technique can be divided into structure-based and ligand-based CADD. Structure-based CADD depends on the knowledge of the target protein to calculate interaction energy for all tested compounds. It is usually used when high-resolution of structural data of the target protein is available. It includes methods such as molecular docking, de novo design, molecular dynamics, and pharmacophore modelling. Ligand-based CADD depends on the knowledge of known active and inactive molecules to perform chemical similarity searches or construct predictive, quantitative structure-activity relation models. It is used when little or no structural information of the target is available. It includes methods such as QSAR, pharmacophore modelling and ligand-based virtual screening[125].

2.2 Molecular Docking

2.2.1 Definition

Molecular docking is a computational tool which has become essential in drug discovery. The goal of this method is to predict the binding position of a specific molecule, the ligand, in relation to another, usually larger, called the receptor[129]. Docking generates an ensemble of 3D conformers of a complex, using the known structures of its free components. In protein-ligand docking, this entails a search through different ligand conformations and orientations within the target protein which are then ranked by the binding affinity of each alternative[99].

The first step of a docking study is defining the area on the protein where the ligand may bind, the binding region. If the location of the binding site is known, programs usually allow the user to restrict the binding region to a specific section of the protein. If nothing is known about the binding site, a blind docking can be made, in which the entire surface of the target is scanned for putative binding pockets. Predictively, blind docking is much less reliable and should only be used as a last resort[130].

There is a large and ever-increasing number of molecular docking programs. All of them involve the search for the preferred poses of the ligand in relation to the receptor. The two main components for a molecular docking procedure are the search algorithm and the scoring function. The search algorithm generates several possible conformations and orientations of the ligand, and eventually the protein, which fit the ligand into the binding pocket of the target receptor. The scoring function is responsible to generate a score for the different poses generated by the search algorithm and then for ranking them. The score should represent the thermodynamics of ligand-protein interaction in order to detect the true binding models[128].

Since most molecular docking programs are developed for being applied to large databases, they are developed to be fast. This means there are several simplifications in both the search algorithms and the scoring functions. Despite that, at the end of the docking study, the best-scored solution should correspond to a true binding conformation. If experimental data exists, the best-scored pose should be close to what is observed[128].

2.2.2 Search Algorithms

As previously stated, the search algorithm has the job of generating an ensemble of protein-ligand poses, hopefully featuring the correct one. Generating poses for a ligand-protein complex means exploring all six degrees of translational and rotational freedom, as well as the conformational degrees of freedom of the ligand and the protein. This leads to a number of possible conformations that is too computationally expensive to be searched in an acceptable time frame. Therefore, docking algorithms integrate various approximations to efficiently search the pose space without an unreasonable computational time. These tools can be categorized into rigid-body, flexible-ligand and flexible protein algorithms[130].

2.2.2.1 Rigid-body algorithms

Rigid-body algorithms are the most basic and the ones that sample the conformational space the fastest. These algorithms do not consider the conformity flexibility of both the ligand and the protein, only sampling the 6 degrees of freedom of the rotational and translational space. This provides several limitations considering that protein-ligand complexes are very dynamic and flexible, and these conformational variations are not being considered. This method was used in earlier ligand-protein docking studies, when the computational power was less than what is available today. They are currently used only in protein-protein docking, due to the special complexity of these systems[130].

2.2.2.2 Flexible-ligand algorithms

Flexible-ligand algorithms are currently the most widely used. These consider the protein as a rigid body and the protein as fully flexible. They explore the 6 translational and rotational degrees of freedom of the complex as well as the conformational degrees of freedom of the ligand. Being more computational demanding, these algorithms use several approximations to permit their use in an efficient manner. They can be categorized into three groups: systematic methods, random or stochastic methods and molecular simulation methods[130].

a) Systematic search docking algorithms try to explore all conformational degrees of freedom of the ligand, and they are also divided into three categories: conformational search methods, fragmentation methods and database methods[130].
    - Conformational search methods explore systematically all rotatable bonds by 360º using small, fixed increments to generate all possible conformations. With a higher number of rotatable bonds there is a much higher number of conformations generated, with the final result being impossible with the current computational power. Therefore, several restrains on the ligand bonds are applied to reduce the number of conformers generated[130].
    - Fragmentation search methods split the ligand into several fragments which are then successively docked into the binding site and covalently linked in order to recreate the original ligand. Instead, the ligand can be divided into a core fragment which is docked first with the left behind fragments being added in an approach known as "incremental construction" or "anchor and grow procedure"[130].
    - Database search methods use databases of pre-generated conformational ensembles to include flexibility in the docking process, considering intra and intermolecular distances. Using a small set of constrained distances, different poses of the ligand are determined[130].

b) Random or stochastic algorithms search the ligand conformational space by doing random modifications in its conformation, which are then accepted or rejected by a predefined probability function. Six main types of docking methods use random algorithms: Monte Carlo, Genetic Algorithms, Tabu Search, Particle Swarm Optimization, Differential Evolutionary Algorithms and Evolutionary Gaussians Algorithms[130].
    - Monte Carlo methods dock the ligand inside the binding site using many random translations and rotations, decreasing the probability of becoming trapped in a local minimum. The simple energy minimization functions used in these methods do not need any derivative information and are very efficient in stepping energy barriers, allowing a good sampling of the

conformational space. The generated conformations are evaluated by a Boltzmann probability function[130].

- Genetic algorithms are based on genetics and the theory of biological evolution. It starts with an initial population of several ligand poses (chromosomes) generated randomly. Each pose is represented by an individual. Each individual is defined by a set of genes, which describe the ligand conformation and its translation and orientation in relation to the protein. The full set of these variables is called the genotype and the atomic coordinates of the ligand are the phenotype. Through various generations, or cycles, genetic operators such as mutations, crossovers and migrations are applied to random individuals of the population to explore the conformational space. At the end of each generation, at random, individuals are evaluated with conformations with negative evolution being excluded. The process continues until the population satisfies a predefined fitness function. Various programs use these algorithms including GOLD and AutoDock which were used in this work. Differential Evolutionary algorithms are derived from GA methods[130].
- Tabu Search algorithms move from one pose to another, imposing several restrictions to make sure that previous poses are not revisited. The Root Mean Square Deviation (RMSD) of a new conformation is calculated in relation to a “tabu list” featuring the visited poses and used to accept or reject the new conformation[130].
- Particle Swarm Optimization is a simpler and faster process than GA methods. The population of ligand poses is called a “swarm” and the poses are called “particles”. Each ligand moves within the search space, keeping in its memory the pose with the lowest energy[130].

c) Molecular simulation algorithms include Molecular dynamics simulations and Energy minimization.

- Molecular dynamics simulations are based on the integration of the Newton’s equations of motion. This method is broadly used in many computational studies, however, its application in molecular docking procedures is limited. In fact, this technique is not very effective to explore the conformational space. It also shows problems such as difficulty crossing high-energy rotational barriers. On the other hand, they can include explicit solvation and explore low-energy conformations[130].

- Energy minimization methods are used not as a search technique but rather as a complementary method that refines the ligand poses. It works by looking to the relative minimum closest to the initial pose.

2.2.2.3 Flexible-protein algorithms

Many molecular docking studies showed that the conventional algorithms can give satisfactory results, even when the protein is considered as a rigid entity, (lock and key model of molecular recognition). However, many proteins undergo a range of structural changes upon ligand binding. These range from a local rearrangement of side chains near the binding site to less common backbone movements (induced fit model). In order to address this issue, specialized search algorithms were developed to account for the partial flexibility of the protein[130].

2.2.3 Scoring functions

Scoring functions are responsible for outlining the correct poses from the incorrect ones. In order to achieve its goal in a reasonable time, the binding affinity is estimated by using several assumptions and simplifications[129]. Considering that numerous physical phenomena involved in molecular recognition are not considered, the accuracy can be compromised. Therefore, the development of scoring functions is not easy. Their accuracy can be evaluated by their ability to achieve these goals: (1) It must estimate the interaction between the ligand and the receptor, with this value being proportional to the Gibbs energy of association; (2) The poses of a ligand must be ranked correctly with the best scored being similar to the pose observed experimentally; (3) It must be able to distinguish molecules that bind the target from those that do not, with the ones that bind having a higher score; (4) It must be fast enough so it can be used in molecular docking[130].

The current number of scoring functions is large and always increasing. They can be separated into four main groups: force field scoring functions, empirical scoring functions, knowledge-based potentials and consensus scoring[130].

2.2.3.1 Force Field Based Scoring Function

Force Field Based scoring functions are based on molecular mechanics force fields such as AMBER. Instead of estimating the free energy of binding they estimate the interaction energy between the protein and the ligand. Originally these scoring functions only accounted for non-bonded terms (Van der Waals, electrostatic) but nowadays other terms such and hydrogen bonds, are also taken into account[130].

The force field used were design to model enthalpy gas-phase contributions to structure and energetics. Therefore, important terms for the ligand-receptor interaction, for instance solvation and entropic terms, were not included. This is corrected by the inclusion of additional terms. Implicit solvation methods such as GBSA or PBSA account for the desolvation energies and a torsional entropy term estimates the conformational entropy lost with the binding process[130].

The downside of this scoring function is the fact that it requires the use of cut-off distances for the treatment of non-bonded interactions. These are chosen arbitrarily which compromises the accurate treatment of long-range effects in the binding process[130].

2.2.3.2 Empirical Based Scoring Function

Although the terms in Empirical Based scoring functions have counterparts in force-field molecular mechanics, they are usually simpler. These scoring functions decompose the overall Gibbs energy of association into components (ΔG_i), as it can be seen on equation 7. In this equation w_i are weight factors that are derived from regression analysis on a training set of protein-ligand complexes with experimentally known binding affinities[130].

$$Score = \sum_i w_i \ \Delta G_i$$

Equation 7 - General Empirical based scoring function.

However, the usage of experimental data means that we cannot be secure that these functions will be able to predict the binding affinity of ligands which are very different from the ones used in the training set. This problem has been declining with the rapid increase in the number of protein-ligand complexes with known 3D structures and affinities, which allows the developing of more general empiric scoring functions[130].

2.2.3.3 Knowledge Based Scoring functions

Knowledge Based scoring functions reproduce experimental data using statistical methods instead of reproducing binding affinities. These functions use statistic potentials to predict the frequency of occurrence of typical interactions, such as different atom-atom pair contacts, obtained from experimentally determined structures. This method assumes that if an interatomic distance is more frequent then average, it represents a favourable contact. The general formula can be seen in equation 9. In this equation *i* and *j* stand for a protein atom and a ligand atom, r is the respective distance, N is the number of all possible atom pairs and u_{ij} corresponds to the pairwise potential between atom i and j[130].

$$Score = \sum_{i,j}^{N} u_{ij}(r)$$

Equation 8- General Knowledge based scoring function.

The main advantage of knowledge-based scoring function is its computational simplicity, only requiring knowledge of a set of protein-ligand complex structures, which has been increasing over the years. It also is as fast as empirical scoring functions. The main disadvantage is that their parameterisation is limited by the sets of complex structures known which are used to develop the algorithm[130].

2.2.3.4 Consensus scoring functions

Consensus scoring functions use information from different scoring functions in order to improve the probability of finding the correct solution. Each scoring function is able to predict the pose but cannot predict the binding affinity since the terms used to describe this interaction are incomplete. There have been various studies which demonstrated that using consensus scoring functions can improve the performance by compensating the deficiencies of each scoring function, being able to reduce the number of false positives identified.[130]

2.2.4 Software

2.2.4.1 Autodock 4

Autodock was originally developed by Morris and co-workers and released 1990. Its latest version, Autodock 4, was released in 2009. This free protein-ligand docking program is one of the most common and most cited software in this field. It has shown good accuracy and high versatility, which makes this program very appealing, mainly for beginners[99,131].

To search the conformational space around the protein, Autodock uses a grid-based method in which a grid is placed on the protein and a probe atom is sequentially placed at each grid point. The interaction energy between the probe and the target is calculated and stored in the grid that serves as a lookup table during the docking simulation. This program primarily uses a Lamarckian genetic searching algorithm, although Monte Carlo simulated annealing and a traditional genetic algorithm are also available[131].

As previously stated, Autodock uses a Lamarckian search algorithm. In the more common Darwinian genetic search algorithms, there is a one-way transfer of information, from the genotype to the phenotype. On the other hand, the Lamarckian algorithm allows each individual conformation to search their conformational space, find their local minima, and pass this information to the next generation. This is an inverse mapping function, in which a genotype is acquired from a given phenotype[131,132].

Autodock4 uses an empirical scoring function for estimating the Gibbs energy of the ligand-protein association process ($\Delta_a G$,) in an aqueous environment. The goal is to capture the complex enthalpic and entropic contributions in a simplified way. This process is then described by a hypothetical mechanism involving two steps, as seen in figure 6.

$$\begin{tikzcd}
\boldsymbol{P + L} \arrow[d, "(\mathbf{1})"'] \\
\boldsymbol{P_{bound} + L_{bound}} \arrow[d, "(\mathbf{2})"'] \\
\boldsymbol{PL}
\end{tikzcd}$$

Figure 6 - Hypothetical mechanism for a protein-ligand association.

In the first step, the molecular fragments (**P** and **L**) are rearranged, assuming the geometries (***P***$\boldsymbol{P_{bound}}$ and $\boldsymbol{L_{bound}}$) adopted in the complex. In the second step, the rearranged species associate with each other, maintaining their geometries and originating the complex (**PL**). The scoring function is then calculated according to this approach.

$$Score \approx \Delta_a G \Leftrightarrow Score \approx \Delta_a G_1 + \Delta_a G_2$$

Equation 9 - Calculation of the Autodock 4 scoring function according to the hypothetical mechanism for a protein-ligand association presented in figure 6.

In equation 9, $\Delta_a G_1$ and $\Delta_a G_2$ are the Gibbs energies associated with the steps 1 and 2, respectively. However, the above-mentioned scoring function (*Score*) is only used for comparative purposes. This means that components considered as non-discriminative are neglected. These include the significant entropy reduction associated with the transformation of the six rotational and vibrational modes of the separated species (protein and ligand) and the six vibrational normal modes in the protein-ligand complex. In this context, $\Delta_a G^o{}_1$ is considered to have an energetic nature:

$$\Delta_a G_1^o \approx \Delta V_1 \Leftrightarrow \Delta_a G_1^o \approx \Delta E_{rearr}(L) + \Delta E_{rearr}(P) \Leftrightarrow$$
$$\Delta_a G_1^o \approx E(L_{bound}) - E(L) + E(P_{bound}) - E(P)$$

Equation 10 - First component of the Autodock 4 scoring function, associated with the step 1 of the hypothetical mechanism for a protein-ligand association process presented in figure 6.

In this equation E(X) represents the energy of a molecular specie X (X = L, P, P_{bound} or L_{bound}). If a rigid-protein search algorithm is adopted, E(P_{bound}) = E(P) and the second term of equation 10 is neglected.

The second component ($\Delta_a G^o{}_2$) is considered to have mix energetic/entropic nature. This component uses pairwise terms to assess the protein-ligand interactions, an empirical method to estimate the contribution of the desolvation process and a corrective term for estimating the decrease of torsional entropy associated with side chains/groups involved in the binding process[132].

$$\Delta_a G_2 = \Delta_a E_{2,\mathrm{L\text{-}J}} + \Delta_a E_{2,\mathrm{hbond}} + \Delta_a E_{2,elec} + \Delta_a G_{2-desolv} + \Delta_a S_{2,tors} \Leftrightarrow$$

$$\Delta_a G_2 = w_{vdW} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + w_{hbond} \sum_{i,j} E(\theta_{ij}) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) +$$
$$w_{el} \sum_{i,j} \frac{q_i \; q_j}{\varepsilon(r_{ij}) r_{ij}} + w_{dessolv} \sum_{i,j} (S_i \; V_j + S_j \; V_i) \; e^{-r_{ij}^2/(2\sigma^2)} + w_{tors} \Delta N_{tors}$$

Equation 11 - Second component of the Autodock 4 scoring function, associated with the step 3 of the hypothetical mechanism for a protein-ligand association process presented in figure 6.

The weight constants w were optimized, using a least squares fitting using a set of Gibbs energies of association determined by experimental techniques. The first term ($\Delta_a E_{2,L-J}$) is a typical Lennard-Jones potential energy for estimating the dispersion/short-range repulsion interactions. The associated parameters (A_{ij} and B_{ij}) were obtained from the Amber force field. The second term ($\Delta_a E_{2,hbond}$) is a directional hydrogen-bond potential energy. The associated parameters (C_{ij} and D_{ij}) are fitted for obtaining a maximal well depth of 5 kcal·mol^{-1} at 1.9 Å for the OH and NH hydrogen-bonds, and a depth of 1 kcal·mol^{-1} at 2.5 Å for the SH ones. $E(\theta_{ij})$ is a penalty energy factor, which increases with the deviation on the hydrogen-bond angle θ_i away from its ideal value. The third term ($\Delta_a E_{2,elec}$) is a Coulomb potential energy used for estimating the electrostatic interactions. The fourth term ($\Delta_a G_{2,desolv}^o$) is associated with the desolvation process. This term has a mixed energetic/entropic nature. It is based on the volume (V) of the atoms surrounding a given atom, weighted by a solvation parameter (S) and an exponential term based on the distance. The distance weighting factor σ was set to 3,5 Å. The final term ($\Delta_a S_{2,tors}$) is associated with the loss of torsional entropy upon binding. This term is directly proportional to the variation on the number of rotatable bonds (ΔN_{tors}), from intermediate unbound state (P_{bound} + L_{bound}) to the final bound state (PL). The remaining terms have the following physical meanings: r_{ij} is the distance between the atoms *i* and *j*, q_i and q_j are the respective charges and $\varepsilon(r_{ij})$ is a distance-dependent dielectric constant[133].

2.2.4.2 Autodock Vina

Following the success of previous Autodock versions, Autodock Vina was developed by Trott and Olson and released in 2009. Vina maintains some of the original ideas from Autodock 4, but it is conceptually different. It has shown to be faster and that it predicts binding poses more accurately then Autodock 4[99,134].

Vina uses an iterated local search global optimizer, in which several steps consisting of a mutation and a local optimization are taken, each being accepted according to the Metropolis criterion. The local optimization is accomplished using the Broyden-Fletcher-Goldfarb-Shanno (BFSG) method. BFSG uses both the value of the scoring function and its derivatives. Although it may take longer to evaluate the derivatives, it can speed up the optimization considerably. The number of steps necessary for obtaining good solutions depends on each application[134].

The scoring function (c) used by Vina, is a combination of the knowledge-based and empirical scoring functions. It is expressed by a pairwise additive potential, as is shown on equation 12. The respective summation includes all pairs of atoms (i,j) which can change the respective relative distance (r_{ij}) during the molecular docking procedure, with the exception of those involved in 1-4 interactions[134].

$$c = \sum_{i<j} f_{t_i t_j}(r_{ij})$$

Equation 12 - Autodock Vina scoring function.

In equation 12, t_i and t_j are the types assigned to atoms *i* and *j* respectively. In this equation, $f_{t_i t_j}$ is an effective interaction potential between these atoms. This scoring function can then be decomposed in an intramolecular (*c_{intra}*) and an intermolecular component (*c_{inter}*):

$$c = c_{intra} + c_{inter}$$

Equation 13 - Decomposition of the Autodock Vina scoring function.

The search algorithm attempts to find the global minimum of c, which corresponds to a conformation designated as k. Its Gibbs energy of association ($\Delta_a G_k$) is estimated from the correspondent intermolecular scoring function (c_{inter_k}):

$$\Delta_a G_k \approx g(c_{inter_k}) \Leftrightarrow \Delta_a G_k \approx g(c_k - c_{intra_k})$$

Equation 14 - Estimation of the Gibbs energy of association, for the lowest-scoring conformation obtained with the Autodock Vina search algorithm.

In equation 14, g is an arbitrary strictly increasing smooth function[134].

2.2.4.3 GOLD

Genetic Optimization for Ligand Docking, or GOLD, is a docking program that was originally developed in 1997. This software, resulted from a collaborative project involving the University of Sheffield, the GlaxoSmithKline and the Cambridge Crystallographic Data Centre. This is one of the most cited docking programs in the literature. It uses a genetic algorithm to generate the ligand poses and has an option of four scoring functions to use. These are the Astex Statistical Potential (ASP), ChemPLP, CHEMSCORE, and GOLDSCORE. The docking procedure can be set-up using the Hermes graphical user interface[99,135].

a) The Astex Statistical Potential is an atom-atom potential drawn from a database of protein-ligand complexes. It is a knowledge-based scoring function. Therefore, it generates statistical potentials using information about the frequency of interaction between ligand and protein atoms from existing ligand-protein structures. ASP differs from other statistical potentials by using a reference state, which determinates how the raw distribution of observations is transformed into potentials. The reference state is the expected number of contacts, if there are no interactions between the atoms[136].
b) ChemPLP is an empirical scoring function which uses the Piecewise Linear Potential (PLP) to model the steric complementarily between the protein and the ligand. In the PLP scoring function two different piecewise functions are defined, one for repulsive/attractive, plp, and one for entirely repulsive interactions, rep.

$$f_{plp} = \sum_{p\epsilon\, P_{prot-lig-plp}} plp(p_r, p_A, p_B, p_C, p_D, p_E, p_F) + \sum_{p\epsilon\, P_{prot-lig-rep}} rep(p_r, p_A, p_B\, p_C, p_D)$$

Equation 15 - Piecewise Linear Potential scoring function.

In equation 15, $P_{prot-lig-plp}$ and $P_{prot-lig-rep}$are sets of protein-ligand atom pairs used for the evaluation of each function. The distance between a ligand a protein atom is given by p_r. The parameters p_A to p_F are dependent on the interaction potential chosen[137].

It is the default scoring function for the current version of GOLD. Unlike the PLP scoring function on which it is based, ChemPLP uses some terms from the CHEMSCORE scoring function (see below). These include the distance and angle-dependent terms associated with hydrogen and metal bonds. The intraligand interactions are estimated using the Tripos force field and a heavy-atom clash term. The ChemPLP function is presented in equation 16. In this equation, f_{plp} stands for the piece linear potential, f_{hb} for the hydrogen bonds, f_{met} for the metal interactions, f_{clash} for the ligand clash potential, f_{tors} for the ligand torsional potential and c_{site} for a quadratic potential responsible to guide the calculations to the binding site[137].

$$f_{CHEMPLP} = f_{plp} + f_{hb} + f_{hb-ch} + f_{hb-CHO} + f_{met} + f_{met-coord} + f_{met-ch} + f_{met-coord-ch} + f_{clash} + f_{tors} + c_{site}$$

Equation 16 - ChemPLP scoring function.

c) CHEMSCORE is an empirical scoring function derived from a set of 82 protein-complexes with experimentally measured binding affinities. It estimates the free Gibbs energy of association using equation 17[138,139].

$$ChemScore \approx \Delta_a G_k \Leftrightarrow ChemScore = \Delta_a G_{ref} + \Delta_a G(L) + \Delta_a G(PL)$$

Equation 17 – General decomposition of the CHEMSCORE scoring function.

In equation 17 $\Delta_a G_{ref}$ is a reference value, $\Delta_a G(L)$ is the component associated with the ligand's conformational rearrangement upon binding and $\Delta_a G(PL)$ is the component associated with the protein-ligand interactions.

The ligand conformation rearrangement component $\Delta_a G(L)$ is considered to have pure energetic nature. It reflects the energy penalty associated with the adoption of a non-optimized geometry by the ligand, in the PL complex, for maximizing its interactions with the protein target[138] (see equation 18).

$$\mathit{\Delta}_a S(L) \approx 0 \Rightarrow \mathit{\Delta}_a G(L) \approx \mathit{\Delta}_a E(L) \Leftrightarrow \mathit{\Delta}_a G(L) = E_{PL}(L) - E(L)$$

Equation 18 - The ligand conformation rearrangement component of CHEMSCORE scoring function.

In the previous equation, $\mathit{\Delta}_a S(L)$ is the ligand conformation rearrangement entropy, $\mathit{\Delta}_a E(L)$ the ligand conformation rearrangement energy, $E_{PL}(L)$ the energy of the ligand in the geometry adopted upon binding with the protein, and $E(L)$ the energy of the ligand in its optimised geometry.

On the other hand, the protein-ligand component $\mathit{\Delta}_a G(PL)$ is calculated according to equation 19[138]:

$$\mathit{\Delta}_a G(PL) = \mathit{\Delta}_a G_{hbond}(PL) + \mathit{\Delta}_a G_{Ma}(PL) + \mathit{\Delta}_a G_{lipo}(PL) + \mathit{\Delta}_a S_{rot}(PL) + E_{clash}(PL) + E_{cov}(PL)$$

Equation 19 - Decomposition of the protein-ligand component of CHEMSCORE scoring function.

In this equation there are a mix of entropic-energetic, pure entropic and pure energetic terms. The entropic-energetic terms are the hydrogen-bond $\mathit{\Delta}_a G_{hbond}((PL)$, the metal-acceptor $\mathit{\Delta}_a G_{Ma}(PL)$ and the lipophilic $\mathit{\Delta}_a G_{lipo}((PL)$ terms. The rotameter term $\mathit{\Delta}_a S_{rot}((PL)$ is the only that is pure entropic. The pure energetic terms are the clash $E_{clash}(PL)$ and the covalent $E_{cov}(PL)$ energies. The mix entropic-energetic terms can be calculated by the following general equation[138]:

$$\mathit{\Delta}_a G_x(PL) = \mathit{\Delta}_a G_{X.opt}(PL) \sum_{i=1}^{n_X} f_i;\ X = hbond, Ma \text{ or } lipo \text{ and } 0 \le f_i \le 1$$

Equation 20 - General equation for calculating the mixed entropic-energetic terms associated with the protein-ligand component of CHEMSCORE scoring function.

In equation 20, n_X is the number of atomic pairs (one atom belonging to the ligand and other to the protein) associated with an interaction of the type X, f_i is the effectiveness factor of the i-th of these interactions and $\mathit{\Delta}_a G_{X.opt}(PL)$ is the Gibbs energy for an optimal interaction (f = 1) of this type.

The rotamer term $\Delta_a S_{rot}(PL)$ represents the entropy penalty, associated with the rotamers of ligand that are constrained due to interactions with the protein. This term is calculated in a similar way to those used in equation 20.

$$\Delta_a S_{rot}(PL) = \Delta_a S_{rot,max}(PL) \sum_{i=1}^{n_{rot}} f_i\,;\ 0 \leq f_i \leq 1$$

Equation 21 - Equation for calculating the rotamer term associated with the protein-ligand component of CHEMSCORE scoring function.

In equation 21, n_{rot} is the number of ligand rotamers that are constrained upon binding, f_i is the effectiveness factor of the *i*-th of these rotamers and $\Delta_a S_{rot,max}(PL)$ is the maximum entropy penalty (correspondent to f = 1) associated with a rotamer of this type[138].

The clash energy term $(E_{clash}(PL))$ is associated with the repulsive interactions involving atomic pairs dominant at short distances ($r \leq r_{clash}$). This term is calculated by equation 22[138].

$$E_{clash} = \sum_{i=1}^{n_{clash}} \varepsilon_i\,(r_i, r_{clash_i})$$

Equation 22 - Equation for calculating the clash energetic term associated with the protein-ligand component of CHEMSCORE scoring function.

In equation 22, n_{clash} is the number of heavy atomic pairs (one atom belonging to ligand and the other to protein), that are close in contact and $\varepsilon_i(r_i, r_{clash_i})$ is the clash energy of the *i*-th of these pairs that is characterized by a distance r_i and a clash distance r_{clash_i}. This energetic quantity can be calculated by equation 23[138].

$$\varepsilon_i(r_i, r_{clash_i}) = \begin{cases} 0; \ r_i > r_{clash_i} \\ \dfrac{20}{\Delta_a G^o_{hbond,opt}} \dfrac{(r_{clash_i} - r_i)}{r_{clash_i}}; \ r_i \leq r_{clash_i} \text{ and the atomic pair } i \text{ is involved in a hydrogen-bond.} \\ \dfrac{20}{\Delta_a G^o_{Ma,opt}} \dfrac{(r_{clash_i} - r_i)}{r_{clash_i}}; \ r_i \leq r_{clash_i} \text{ and the atomic pair } i \text{ is involved in a metal-acceptor interaction.} \\ 1 + \dfrac{4 \ (r_{clash_i} - r_i)}{r_{clash_i}}; \ r_i \leq r_{clash_i} \text{ and the atomic pair } i \text{ is not involved in any of the previous interactions.} \end{cases}$$

Equation 23 - Clash energy for an atomic pair, characterized by a distance r_i and a clash distance r_{clash}.

The covalent energy term ($E_{cov}(PL)$) is associated with the covalent bonds eventually established between the ligand and the protein. This term can be calculated using equation 24[138].

$$E_{\text{cov}} = \sum_{i=1}^{n_{tc}} \varepsilon_{\text{tors}}(\omega_i) + C_{\text{cov}} \sum_{j=1}^{n_{ac}} K_j \, (\theta_j - \theta_{o,j})^2$$

Equation 24 - Equation for calculating the covalent term associated with the protein-ligand component of CHEMSCORE scoring function.

In equation 24, the first summation is over all n_{tc} dihedral angles involved in the covalent linkage and the second one is extended to all n_{ac} covalent bond angles around the same linkage. In this equation, $\varepsilon_{tors}(\omega_i)$ is the torsional energy associated with the dihedral angle (ω_i), K_j is the force constant of the bond angle number, j of magnitude θ_j, $\theta_{0,j}$ the ideal magnitude for this angle and C_{cov} a constant used to balance the covalent bond term against the rest of the CHEMSCORE scoring function.

d) The GOLDSCORE function is the original scoring function used by GOLD. This is a force field scoring function, which is used for estimating the association energy ($\Delta_a E$) according to the two-step hypothetical mechanism presented in figure 6. Therefore, this scoring function can be calculated using equation 25[138].

$$GoldScore \approx \Delta_a E \Leftrightarrow GoldScore \approx \Delta_a E_1 + \Delta_a E_2 \text{ (25a)}$$

$$\Delta_a E_1 = \Delta E_{rearr}(L) + \Delta E_{rearr}(P) \Leftrightarrow$$

$$\Delta_a E_1 \approx E(L_{bound}) - E(L) + E(P_{bound}) - E(P) \text{ (25b)}$$

$$\Delta_a E_2 = \Delta_a E_{2,\text{L-J}} + \Delta_a E_{2,\text{hbond}} \quad \text{(25c)}$$

Equation 25 - General formulation for the GOLDSCORE scoring function.

In equation 25, the different terms have similar physical-meanings to the correspondent quantities described for the AutoDock 4 scoring function (figure 6 and equations 8 to 10). If a rigid-protein search algorithm is adopted, E(P$_{\text{bound}}$) = E(P) and second term of equation 25b is neglected.

2.2.4.4 LeDock

LeDock is flexible small-molecule docking software developed by Hongtao Zhao and co-workers[140].

LeDock combines a genetic algorithm with simulated annealing search to generate the first generation of docking poses. The conformation of the ligand is randomly changed at the start of each simulated annealing search, so that each search starts with a different pose. This software uses knowledge based scoring function which can be calculated by the following equation[140].

$$LeDockScore \approx \Delta_a G^o \Leftrightarrow$$

$$LeDockScore = \alpha \sum_{i=1}^{n_L} (E_i^{LJ} + E_i^{hb}) \times H\left(\left|E_i^{LJ} + E_i^{hb}\right| - E_{cut}\right) + \beta(r) \sum_{i=1}^{n_L} E_i^{el} + \gamma\ E_L^{str}$$

Equation 26 - LeDock scoring function.

In this equation, the summations are extended to all the n_L atoms of the ligand. Each of their terms represents a specific interaction (Lennard-Jones + hydrogen bond in the first summation and electrostatic in the second summation) between an atom *i* of the ligand with all atoms of the protein. The strained energy E_L^{str} has a similar physical meaning than that of the ligand rearrangement energy ($\Delta E_{rearr}(L)$) term used in equation 24b. In the first summation, H is the Heaviside step function (see equation 27) and E_{cut} is the cut-off energy for Lennard-Jones + hydrogen bond interactions.

$$H(x) = \begin{cases} 0; & x < 0 \\ 1; & x \geq 0 \end{cases}$$

Equation 27 - The Heaviside step function.

As E_{cut} is a positive value, $H\left(\left|E_i^{LJ} + E_i^{hb}\right| - E_{cut}\right)$ prevents the docking algorithm of calculating negligible interactions of this type. The coefficients α, $\beta(r)$ and γ are fitted, using a least squares procedure, for reproducing experimental values of $\Delta_a G^o$ obtained for a large number of protein-ligand complexes. In particular, β(*r*) is a distance dependent function, which accounts for both electrostatic screening and desolvation. The Lennard-Jones (E_i^{LJ}), hydrogen bond (E_i^{hb}) and electrostatic (E_i^{el}) interactions of the ligand atom *i* with the protein are calculated respectively as:

$$E_i^{LJ} = \sum_{j=1}^{n_P} 4\varepsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right) \text{(28a)}$$

$$E_i^{hb} = \sum_{j=1}^{n_P} w_{ij}(r_{ij} - r_{cut})\ H(r_{cut} - r_{ij}) \text{ (28b)}$$

$$E_i^{el} = \sum_{j=1}^{n_P} \frac{q_i\, q_j}{r_{ij}} \text{ (28c)}$$

Equation 28 - The Lennard-Jones (E_i^{LJ}), hydrogen bond (E_i^{hb}) and electrostatic (E_i^{el}) interactions of a ligand atom i with the protein.

In general, for these equations, r_{ij} is the distance between the ligand's atom *i* and the protein's atom *j*. In equation 28a, σ_{ij} is the distance for which the Lennard-Jones interaction energy between atoms *i* and *j* is null and ε_{ij} is the symmetrical of the minimum value for this interaction energy. In equation 28b $w_{ij}(r_{ij} - r_{cut})$ is the energy of the hydrogen-bond that depends on the nature of the atoms involved and on the distance, while r_{cut} is the cut-off distance (minimum distance for a non-null hydrogen bond interaction). $H(r_{cut} - r_{ij})$ is the Heaviside step function that imposes this constrain. In equation 28c q_i and q_j are the charges of atoms *i* and *j* respectively.

2.3 Virtual Screening

2.3.1 Introduction

Virtual screening (VS) is a computational technique which includes several methodologies usually divided in two major groups: ligand-based virtual screening and receptor-based virtual screening.[98]

Ligand-based virtual screening methods try to find molecules with similar physical and chemical properties, based on the belief that similar compounds will have similar effects on a drug target. These methods discard all information related to the drug target only considering the ligand. The main disadvantage is that these methods need a significant amount of activity data for the compounds that are studied in order to get reasonable results[128].

Receptor-based virtual screening methods, also known as structure-based methods, need a 3D structure of the target in order to perform multiple molecular docking studies. The scores obtained are then used to distinguish the ligands that bind strongly to the target from those that do not. These methods usually are more reliable and accurate than ligand-based methods. Adding to that, the increased amount of available 3D structures means that receptor-based methods are gaining significant importance over ligand-based methods[128].

Similar to high-throughput screenings, VS methods are used in the beginning of the drug discovery process with the purpose of enriching the initial library with active compounds. The main advantage that VS has over HTS is that using VS it is possible to evaluate thousands of compounds in a matter of hours and reduce the number of compounds that have to be synthesized or purchased, therefore decreasing the overall cost[98].

2.3.2 Validation

Before performing the Virtual Screening to a large database, it is important to make sure that the computational results can be trusted. Therefore it is important to validate the molecular docking protocol and the VS procedure[128].

One way to validate the molecular docking protocol is to assess the quality of the docked poses. This can be done by redocking experiments. These experiments consist of docking ligands for which the experimental binding modes have already been determined. The standard way to compare the redocked pose to the experimental one is to calculate the Root Mean Square Deviation between them. Cut-off values used to classify poses as correct are usually around 2 Å[128].

The more accurate way to evaluate the performance of a VS procedure is to quantify its ability to discriminate between active and inactive molecules. To evaluate this ability, a database of active compounds and decoys is created. Active molecules are usually found in databases such as ChEMBL[141] or in the literature. Since it is rare to find information regarding inactive molecules, random molecules are often used as decoys. This assumes that these molecules will not be active for the target in study[128]. These molecules can be obtained using the Database of Useful Decoys – Enhanced (DUD-E)[142]. For each active molecule, this server can generate 50 decoys with similar physico-chemical properties but with a different 2D topology. This means that, with only 25 active molecules, 1250 decoys are obtained, and the full database used for validation will have 1275 compounds[142].

In a perfect VS protocol, the worst active molecule would have a better score than the best score for an inactive molecule. However, there is a significant overlap between the scores of the active and inactive molecules. Most statistical tools used to quantify the performance of the VS protocol are based on a threshold value to classify molecules in active or inactive. Active molecules scored above the threshold are called true positives while decoys are called false positives. Decoys under the threshold are called false positives and actives are false negatives[128].

Different metrics have appeared to evaluate the performance of VS protocols including receiver operating characteristics (ROC) curves, area under the curve (AUC), enrichment factor (EF), robust initial enhancement (RIE), Boltzmann enhanced discrimination of ROC (BEDROC), predictive curves (PC) and total gain (TG)[128,143].

The ROC curve was created to graphically represent the performance of a ranking method, not considering any threshold, and therefore giving a general view of the performance. It is a plot of the true positive rate (TPR) in function of the false positive rate (FPR). The TPR and FPR are the true positives and false positives expressed as a percentage of the total number of actives and decoys. A random ranking method would lead to a plot line with a slope of 1. Higher initial slopes indicate a better performing VS protocol. ROC curves are analysed by calculating the area under the curve which is the probability of ranking actives above decoys. A random predictor will display an area under the ROC curve of 0.5, while a AUC of 1 would correspond to the perfect scenario[125,128].

The predictive curves quantify the predictive power of the scoring functions. The total gain is bounded by 0 and 1 and it is calculated from the PC curves. TG quantifies the discrimination of actives over decoys attributable to score variations. High TG values, over 0.4, combined with an AUC over 0.5 indicate a good performance from the VS protocol[143].

Since the goal of a VS process is to generate a list of compounds to test experimentally, it is only practical to test the better few hundred ranked molecules of the initial database. Therefore, it is important to use metrics that evaluate early recognition of actives. One way to evaluate early recognition is by calculating the logarithmic curve of the ROC which will gave a greater emphasis to the early performance[128].

Other way to evaluate early recognition is by using the enrichment factor. EF is a measure of how much the sample is enriched with actives at a given threshold. It is calculated by the ratio between the fraction of active compounds recovered by the fraction of the screened library at the chosen threshold, as seen on equation 29. In order to display a more general view of the performance, it is usual to report the EF for more than 1 threshold. Typical EF are reported at 1%, but it is also common to report EF up to 20%[128].

$$EF = \frac{Nº\ actives\ recovered}{Nº\ actives} x \frac{Nº\ total}{Nº\ screened}$$

Equation 29 - Calculation of the Enrichment Factor.

The EF has the disadvantage of being strongly dependent of the number of actives, and because of its lack of discrimination before the threshold. The robust initial enhancement is an early recognition metric which addresses the disadvantages of EF. Unlike EF, RIE includes contributions from all actives into the final score. It distinguishes the situation where all actives are ranked at the beginning from the situation where all actives are ranked closed to the threshold. RIE uses a decreasing exponential weight as a function of the actives ranks. The general equation of RIE is shown on equation 30 in which *n* is the number of actives, *N* is the number of compounds, x_i is the relative scaled rank and α is the weight parameter. While RIE addresses the shortcoming in EF, it lacks the advantages of ROC[143,144].

$$RIE = \frac{\sum_{i=1}^{n} e^{-\alpha x_i}}{\frac{n}{N}\left(\frac{1 - e^{-\alpha}}{e^{\alpha/N} - 1}\right)}$$

Equation 30 - General equation of the calculation of RIE.

The Boltzmann enhanced discrimination of ROC is a normalization of the RIE bounded by 0 and 1 and can be calculated using equation 31. BEDROC contains the discrimination power of RIE and the statistical significance from ROC as well as its well-behaved boundaries[143,144].

$$BEDROC = \frac{RIE - RIE_{min}}{RIE_{max} - RIE_{min}}$$

Equation 31 - Calculation of BEDROC.

2.4 Molecular Dynamics simulations

2.4.1 Introduction

Molecular dynamics (MD) simulations are a computational technique that simulates the dynamic behaviour of a flexible molecular system in function of time. This method is based on numerical integration of the Newton's motion equations for a molecular system:[145]

$$m_i \frac{\partial^2 r_i(t)}{\partial t^2} = -\frac{\partial E(r^N)}{\partial r_i};\ i = 1,\ 2,\ \dots\ , N$$

Equation 32 - The Newton's motion equation for a molecular system.

In equation 32, r^N = (r_1, r_2, …, r_n) represents the coordinates of the N atoms composing the molecular system, $r_i(t)$ the coordinates of the atom *i* as a function of time *t*, m_i the respective atomic mass and $E(r^N)$ the potential energy associated with that system[145].

The MD method was originally developed to simulate simple systems, with its first application being in 1957. The first simulation of a biomolecule was performed in 1976, featured a Bovine Pancreatic Trypsin Inhibitor with 58 residues and covered only 9.2 ps of simulation time[146]. MD simulations are now widely used as a tool to investigate the structure and dynamics of proteins in various conditions such as studies of ligand binding and protein re-folding[147].

2.4.2 Ensembles and time step

An MD trajectory only provides information of atomic positions, velocities and single-point energies. To obtain the general view of the system, statistical mechanics are used. These rigorous mathematical expressions relate the distributions and motions of atoms and molecules to the macroscopic properties of the system. Using this information, it is possible to predict changes in the binding free energy of a particular drug candidate or the mechanisms of a conformational change in a particular protein[146].

The macroscopic state of a system can be characterized with a small number of properties, called state functions. There are several state functions with some being temperature, pressure, and the number of particles. To define the thermodynamic state of the system, the value of α+2 state functions is necessary, with α being the number of components on the system. Each macrostate corresponds a to very large number of microstates. Each microstate can be characterized by the positions and moments of each particle of the system. A set of microstates with the same thermodynamic restrictions is called an ensemble[145].

With different sets of α+2 defined state functions, there are different ensembles. The most common are: the canonical ensemble (*NVT*), where the number of particles (*N*), the volume (*V*) and the temperature (*T*) are constant; the microcanonical ensemble, (*NVE*), where number of particles, the volume, and the total energy (*E*) are constant; the isothermal-isobaric ensemble (*NPT*), where the number of particles, the temperature and the pressure (*P*) are constant; and the grand canonical ensemble *(μVT*) where the chemical potential (*μ*), the volume and the temperature are constant. The appropriate ensemble used for simulating a molecular system, depends on the experimental condition desired to replicate[145].

Molecular dynamics is a deterministic method. This means that the state of any system at any future time can be predicted from its current state. Atomic positions and moments are determined by integrating the above-mentioned Newton's equations of motion (see equation 32). In MD simulations, the trajectory of each particle is calculated in order to get a set of chronologically ordered microstates. To achieve this, the resultant force acting in each atom is calculated and the equations of movement are numerically integrated. In order to perform this integration, a time step is necessary[145].

The choice of an appropriate time step is essential. If it is too short a lot of computational power and time will be needed to simulate a short period of time. If the time step is too large, there will be inaccuracies on the integration of the equations. The first MD simulations used very simple potentials such as the hard-sphere potential in which particles move in straight lines at constant velocity between collisions. While this model provided useful results in the past, is not ideal for the simulation of atomic or molecular systems. Potentials currently used have a more continuous nature, for example, in the Lennard-Jones potential the force between two atoms changes continuously as they separate. These more realistic potentials need the integration of the equations of motion into a series of shorter time steps[145].

The time step is usually one order of magnitude smaller than the shortest motion in the system. In biomolecular systems, the shortest motion is the vibrations associated to bond-stretching of atoms bonded to hydrogens. This motion has a vibration period in the order of 10 fs. This means that the recommended time step used for MD simulations is usually 1fs. However, these motions rarely affect the study of the properties of the system. The solution for this problem is to restrict bonds involving hydrogen to their equilibrium values, while not affecting all other bonds. With this restriction, the shortest motions are vibrations involving heavy atoms which are 2 to 5 times slower, allowing the usage of a 2 fs time step[145]. The most common way to restrict bonds involving hydrogen atoms is to use the SHAKE algorithm[148].

2.4.3 Molecular Mechanics, Classic Potential Energy and Force Fields

The concepts of “molecular mechanics”, “classic potential energy functions” and “force fields” are frequently misunderstood. In this section, an effort is made for clarifying these concepts.

Molecular mechanics is a methodology for modelling molecular systems. It is based on the following general principles: a) Molecular mechanics is based in a classic approach; b) The electrons are not explicitly treated; c) Atoms are considered to be point particles, characterized by their charge and mass; d) The interactions between the atoms are described by a classic potential energy ($E(r^N)$)[145].

The classic potential energy function, which emerges from the general principle d) of the molecular mechanics formalism, can be expressed as a sum of a bonded component (E_{bond}) and a nonbonded one ($E_{n/bond}$).

$$E = E_{bond} + E_{n/bond}$$

Equation 33 - General decomposition of a classic potential energy function.

The bonded potential energy can be decomposed into stretching (E_{stret}), bending (E_{bend}), and torsional (E_{tors}), as shown in the following equation:

$$E_{bond} = E_{stret} + E_{bend} + E_{tors}$$

Equation 34 - General decomposition of a bonded potential energy.

The nonbonded potential energy can be expressed as a sum of a Lennard-Jones term (E_{LJ}) and an electrostatic one (E_{el}):

$$E_{n/bond} = E_{LJ} + E_{el}$$

Equation 35 - General decomposition of a nonbonded potential energy.

The usage of an appropriate energy function to describe intermolecular and intramolecular interactions is essential to achieve a successful molecular dynamics simulation. The energy functions usually consist of several parameterized terms which are obtained from experimental or quantum mechanical studies of small molecules. The set of functions with the associated set of parameters is called a force field[146].

Various force fields have been developed, with several being specific for the simulation of proteins. Some of these are AMBER, CHARMM and GROMOS[145].

2.4.4 Boundaries

Treating boundaries and boundary effects properly is essential in a MD simulation. Simulated systems frequently have up to 300000 atoms in a cubic or octahedral cell. One of the issues of using these cells is that, on a cubic box, the molecules close to the edges will suffer different interactions to the ones inside the box. These are called boundary effects. To calculate properties without these effects, one could not count the contribution of the molecules near the border. The issue here is that the molecules near the border can account for 40% of the system leaving too few molecules in order to derive the properties[145].

To solve this issue, periodic boundary conditions can be used. In these conditions, the simulation cell is the central point of an infinite cubic network of its copies. The integration of the equations of motion is only done on the central cell while all other cells mimic them. If a particle leaves the box, it is replaced by an image particle that enters from the opposite side. Using these conditions, a simulation can be performed using a relatively small number of particles, in such a way that all particles experience forces as if all were in fluid[145].

Obviously, the replicated movements of the image molecules are not realistic and coherent in a biologic system. It becomes necessary for the central cell to be big enough so that each molecule does not interact in a significant manner with the replica molecules[145].

In an infinite network consisting of an infinite number of particles, it is not possible to calculate the interaction energy. Intermolecular interactions are only approximately calculated when these calculations only include a finite number of neighbouring molecules. This can be achieved using the minimum image convention or by a spherical cut-off. The minimum image convention centres, in each atom, one cell with the same size and shape as the original. Interactions are calculated for each atom inside that cell, with all other being ignored. Alternatively, in the spherical cut-off approach, a sphere with a predetermined radius is used for the same puropose[145].

2.4.5 Long-Range Interactions

Since the interaction energy calculation cannot include all particles and their infinite images, there must be a cut-off at a certain radius. For short-range, Lennard-Jones, interactions, a reasonably small cut-off (10-12 Å) can be used without significant errors. However, for long-range, electrostatic interactions, the cut-off should be significantly bigger. Nevertheless, for molecular systems containing many charged groups, additional procedures have to be adopted for accounting long-range interactions of this type[145].

The issue of long-range interactions can be solved using the Ewald summation method. In an ionic system, this method will make two changes. Each point charge is neutralized at long distances by the introduction of a spherical charge cloud. The cloud's charge density is calculated by a Gaussian function centred at the ion. Adding the point charges and Gaussian charges cancels the electrostatic potentials and makes it short range, allowing it to be treated by simple truncation. The other change is to introduce a second set of Gaussian clouds with an opposite charge to the first to cancel their effect. It is also necessary to calculate the interaction energy that each charge cloud has with itself. This constant will be subtracted from the result of the Ewald summation[145].

This method is controlled by two parameters, the truncation radius and parameter that controls the variation of the Gaussian charge distribution. A higher value means a less dense Gaussian charge and shorter-range electrostatic interactions. This means a shorter cut-off radius can be used. However, the lesser density of the Gaussian charge leads to a higher volume of calculus. The solution is to use the higher truncation radius possible, which is half the size of the cell, and the smaller possible value. That way the contribution of the interactions outside the radius does not have a significant effect on the system[145].

2.4.6 AMBER

2.4.6.1 Force Field

AMBER is a force field used in MD simulation of proteins, nucleic acids, and carbohydrates. The functional form for its potential energy is described as minimalist and expressed by equations 33 to 35 presented above. The bonded terms, present in equation 33 have the following forms[149]:

$$E_{stret} = \sum_{i=1}^{n_b} E_i, \text{ with } E_i = K_i \ (r_i - r_{i,eq})^2 \text{ and } K_i = \frac{\partial^2 E_i}{\partial r_i^2} \text{ (36a)}$$

$$E_{bend} = \sum_{j=1}^{n_{ba}} E_j, \text{with } E_j = K_j \ (\theta_j - \theta_{j,eq})^2 \text{ and } K_j = \frac{\partial^2 E_j}{\partial r_j^2} \text{ (36b)}$$

$$E_{tors} = \sum_{k=1}^{n_t} E_k \text{ with } V_l = \frac{K_k}{2} \ [1 + cos(n_k \omega_k - \gamma_k)] \text{ (36c)}$$

Equation 36 - Bonded terms of AMBER potential energy function.

In equation 36a, the summation is extended to all n_b bonds of the molecular system. In the same equation E_i is the stretching potential energy associated with the *i*-th bond, K_i the associated force constant, r_i the length of this bond and $r_{i,wq}$ the correspondent equilibrium value[149].

In equation 36b, the summation is extended to all n_{ba} bond angles of the molecular system. E_j is the potential energy associated with the *j*-the bond angle K_j the associated force constant, θ_i the amplitude of this angle and $\theta_{j,eq}$ the correspondent equilibrium value[149].

In equation 36c, the summation is extended to all n_t torsional angles of the molecular system. In the same equation E_k is the potential energy associated with the *k-th* torsional angle, ω_k the amplitude of this angle, K_k the maximum for the referred potential energy, γ_k the torsional angle associated with this maximum and n_k the torsional angle multiplicity (number of energy minima associated with it)[149].

For a molecular system with N atoms, the nonbonded terms that occur in equation 35 have the following forms:

$$E_{LJ} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} E_{LJ}(i,j), \text{ with } E_{LJ}(i,j) = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \text{ (37a)}$$

$$E_{el} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} E_{el}(i,j), \text{ with } E_{el}(i,j) = \frac{q_i \times q_j}{4\pi\varepsilon_o r_{ij}} \text{ (37b)}$$

Equation 37 - Nonbonded terms of AMBER potential energy functional.

In equation 37, both Lennard-Jones and electrostatic terms are expressed as pairwise additive summations of interatomic interactions. These summations include all atomic pairs *(i,j),* occurring at a distance r_{ij} in the molecular system, with the exception of those associated with either 1-2 or 1-3 interactions (figure 7)[149].

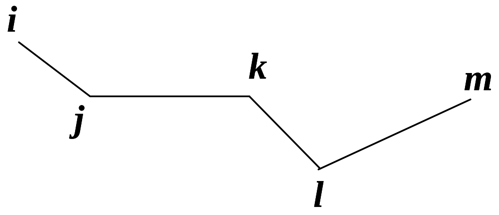


Figure 7 -Classification of the interatomic interactions that occur in a molecular system: i-j *1-2 interaction;* i-k *1-3 interaction;* i-l *1-4 interaction and* i-m *"normal interaction".*

In equation 37a, A_{ij} and B_{ij} are the rearranged Lennard-Jones parameters that can be calculated from the original ones (see equation 37) as:

$$A_{ij} = 4\varepsilon_{ij}\sigma_{ij}^{12}, B_{ij} = 4\varepsilon_{ij}\sigma_{ij}^{6}$$

Equation 38 - Calculation of the rearranged Lennard-Jones parameters (A_{ij} *and* B_{ij}*) from the original ones (*ε_{ij} *and* σ_{ij}*).*

In equation 37b, ε_o is the vacuum permittivity and q_i and q_j are respectively the charges of atom i and j[149].

2.4.6.2 AMBER Package

AMBER is also the collective name for a suite of programs that allows the execution of molecular dynamics simulations. There are three main steps in the MD protocol: system preparation, simulation and trajectory analysis[150].

The main preparation programs are antechamber and LEaP. Antechamber takes a three-dimensional structure and assigns charges, atom types and force field parameters for residues or organic molecules that are not part of standard libraries. LEaP constructs biopolymers from the respective residues, solvates the system and prepares the list of force field terms and parameters. The result of this phase is a file that contains the Cartesian coordinates of all atoms in the system and a file which contains all the other information required, including atom names and masses, force field parameters, lists of bons, angles and dihedral angles[150].

The main simulation program is sander. It uses a replicated data structure, in which each processor is responsible for certain atoms, but all processors know the coordinates of all atoms. During each step, the processors calculate a portion of the potential energy and then a binary tree global communication sums the force vectors. This way each processor has the full force vector components for the atoms it is responsible for. Then the processors perform a MD update step for those atoms and a use another binary tree to communicate the updated positions to all processors[150].

The main analysis programs are ptraj and its successor cpptraj. These programs read the parameter and topology files. They also allow the inspection of the information within that file such as lists of bonds, angles, dihedrals and other[150].

2.5 MM/PB(GB)SA

Scoring functions are efficient, can predict binding modes and distinguish binders from non-binders. However, they are not able to separate between molecules that diverge by less than one order of magnitude. There are other methods to calculate Gibbs energy of association including the thermodynamic perturbation (TP) and thermodynamic integration (TI) methods. These methods are very rigorous and computationally expensive methods, requiring massive sampling of the free ligand and complex in solution. They require the definition of a considerable number of intermediate states between the initial and final states. It is also necessary to perform many independent MD simulations, correspondent to the transition between each pair of consecutive states of this type. Fortunately, there are other methods that only require the end states of the ligand, the receptor, and the complex. They are less expensive than TP or TI formalisms and more accurate than scoring functions. These end-point methods include the Molecular Mechanics/Poisson-Boltzmann Surface Area (MM/PBSA) and the Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) approachs[151].

These methods are very effective to estimate the Gibbs energy of association of small ligands to large biological receptors[152].

In these methods, the Gibbs energy of association of a protein-ligand complex is estimated using the following equation:

$$\Delta_a G = \langle G(PL) \rangle - \langle G(P) \rangle - \langle G(L) \rangle$$

Equation 39 - Estimation of Gibbs energy of association using either the MM/PBSA or MM/GBSA formalisms.

In equation 39, the notation $\langle G(X) \rangle$ stands for the average value of the Gibbs energy of the molecular specie X (with X = PL, P or L). This thermodynamic quantity is then estimated as:

$$\langle G(X) \rangle = \langle E_{bond}(X) \rangle + \langle E_{el}(X) \rangle + \langle E_{LJ}(X) \rangle + \langle G_{pol}(X) \rangle + \langle G_{n/pol}(X) \rangle - T\ S(X)$$

Equation 40 - Estimation of the average value for the Gibbs energy of a molecular specie X, using either the MM/PBSA or MM/GBSA formalisms.

In the last two equations, the different physical quantities within brackets are estimated as ensemble averages for the conformations generated by MD simulatons[100]. In equation 40 all the quantities are referred to the molecular specie X. In this context, $E_{bond}(X)$ is its bonded energy,$E_{el}(X)$ its electrostatic energy and $E_{LJ}(X)$ its Lennard-Jones energy. These quantities are evaluated using an appropriate force field as for example AMBER. In the same equation $G_{pol}(X)$ and $G_{n/pol}(X)$ are respectively the polar and the non-polar contributions to the solvation Gibbs energy of the molecular specie X, T is the absolute temperature and $S(X)$ is the entropy of the same species. The $G_{pol}(X)$ term is obtained by solving the Poisson-Boltzmann (PB) equations for the MM/PBSA formalism or using the generalized Born model (GB) for the MM/GBSA approach. The Poisson-Boltzmann approach leads to very high computational costs, because the PB equation has to be solved every time the conformation of the molecule changes. The Generalized Born method became popular due to its relative simplicity and computational efficiency compared to the PB equations. The $G_{n/pol}(X)$ term is estimated from a linear relation to the solvent accessible surface area (SASA). The entropy $S(X)$ can be estimated by using a normal modes analysis. However, the correspondent $T\ S(X)$ term is often neglected. In fact, this term frequently increases the errors associated with this type of calculations[151,152].

Although the ensemble averages in equation 39 should be estimated from three separate MD simulations (for protein, ligand, and protein-ligand solvated species), it is common to adopt a more simplified procedure. In fact, if the effects associated with the conformation rearrangements of both solvated protein and ligand species are neglected, it is only necessary to perform one MD simulation of the solvated protein-ligand complex[152].

MM/PB(GB)SA calculations can be run using AMBER's MM/PBSA.py python script. In this method representative snapshots from an ensemble of conformations are used to calculate the change in the Gibbs energy between two states[153].

The Solvent Accessible Surface Area (SASA) is a geometric evaluation for the exposure of the amino acids to their environment. The SASA is computationally calculated using a spherical probe, similar to a water molecule, on a full-atom molecular model[154].

2.6 Molecular Visualization Software

2.6.1 PyMOL

PyMOL is a cross-platform molecular graphic system originally released in 2000. Adding to the fact that PyMOL has most of the features seen in most molecular graphics packages, the integrated Python interpreter allows its users to develop an unlimited number of plugins with a wide range of functions. This Python interpreter is the reason for the “Py” portion of the software’s name[155,156].

The most common graphical representations are supported by PyMOL including cartoon ribbons, backbone ribbons, solid surfaces, ball-and-stick, between others. Atoms can be labelled, and dash bonds can be used to illustrate hydrogen bonds and distances between atoms[155].

PyMOL can produce high quality 3D images of both small molecules and larger biological macromolecules. It features an integrated ray-tracing engine capable of converting any displayed view into a publication quality figure. In fact, it is estimated that nearly a quarter of all published images of 3D protein structures were made using PyMOL[155,156].

2.6.2 VMD

Visual Molecular Dynamics was released in 1996 and it designed for the visualization and analysis of biological systems such as proteins and nucleic acids. VMD can display structures using several representations and colouring methods including CPK spheres, licorice bonds, cartoon drawings, and others[157].

VMD was also designed to animate and display the trajectory of a molecular dynamics simulation, imported from files or from a direct connection to a running MD simulation[157].

This software has a graphical user interface as well as a Tcl text interface to allow the users to create their own scripts. VMD was written in C++ and its source code and documentation is available online[157].

2.7 Structural Files

2.7.1 PDB

The Protein Data Bank was created in 1971 by Walter Hamilton in response to a necessity for a central repository for information about biological macromolecular structures. Due to its simplicity and uniformity, the PDB format is still the most popular format for macromolecular structural data[158].

The PDB format features the coordinates of each atom, chemical and biochemical features, experimental data of the structure determination, and some structural features including secondary structure and hydrogen bonds. PDB has defined conventions for naming atoms, residues, and nucleotides[158].

2.7.2 SMILES

The Simplified Molecular Input Line System (SMILES) is an unambiguous and reproducible method for representing small molecules. It was developed in 1985 by David Weininger. SMILLES does not explicitly include hydrogen atoms, with the convention that hydrogens make up the remainder of an atom's normal valence. All other atoms are represented by their atomic symbols surrounded by square brackets, or without square brackets implying the existence of hydrogens. This format also includes information on formal charges, aromaticity, and bonds. Formal charges are included with the number preceded by a + or -. Aromatic atoms are represented with lowercase atomic symbols. Single bonds, double bonds, triple bonds and aromatic bonds are represented by "-", "=", "#" and ":;". Usually there are several SMILES descriptions for the same molecule, all equally valid. This format is used across multiple ligand databases[125].

2.7.3 PDBQT

The PDBQT file was created with AutoDock 4. This file stores atomic coordinates, partial charges and AutoDock atom types necessary for the molecular docking procedure. Both the ligand and receptor PDBQT file require Gasteiger PEOE partial charges and a united-atom representation, including only polar-hydrogens, so that they work with the AutoDock 4 scoring function[159]. This format is also required to perform molecular docking using AutoDock Vina[134].

2.7.4 Open Babel

With the emergence of several file formats to store chemical structure information, there was a need to convert the different formats into the required one for each software. Open Babel was designed to read the many representations of chemical data and it is able to search, convert, analyse, or store molecular data. It can read 82 different formats and write 85, supporting in total 111 chemical file formats[160].

3. Aim of this study

Over the years, the usage of computer-aided drug design as a preliminary stage of drug design has increased. This makes the entire process more cost-efficient and minimizes failures.

Biofilm infections have been recognized as a serious threat to our society. These biological structures are not easily treated by existing antimicrobial treatments and are very hard to remove after their formation. Consequently, it is increasingly important to find new drugs to mitigate biofilm formation.

Quorum sensing is deeply involved on the process of biofilm formation. Impeding this cell-to-cell communication has been shown as a promising way to prevent the formation of biofilms.

Therefore, the aim of the present work was to model promising molecules for blocking quorum sensing and preventing biofilm formation. Different computational methods were used in this study. These include molecular docking, virtual screening, MD simulations and MM/PB(GB)SA calculations. The molecular target in this study is CviR, the quorum sensing receptor from *Chromobacterium violaceum*, an opportunistic pathogen used as a model organism for QS research. The workflow of this work is presented on figure 8.

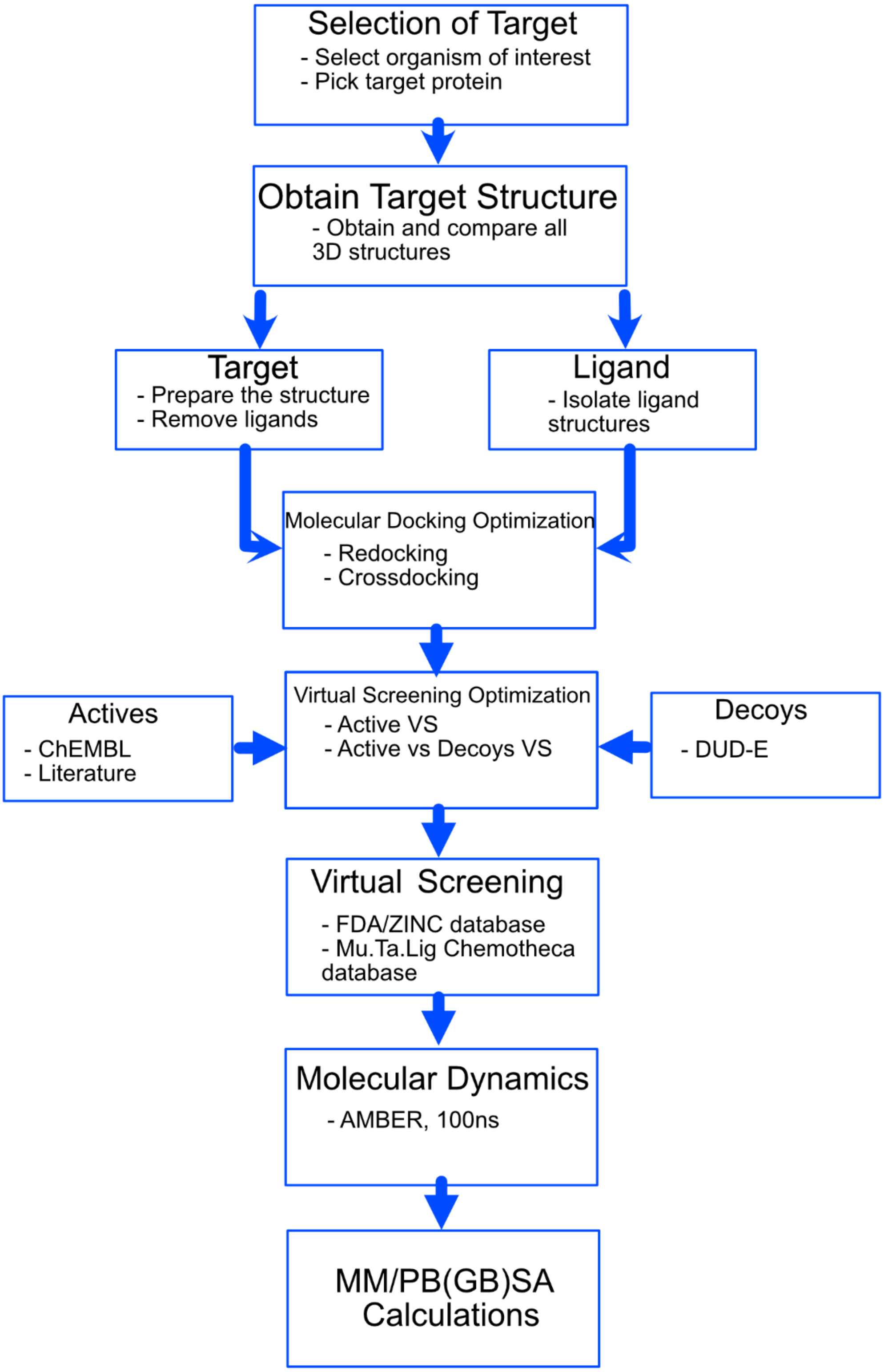


Figure 8 - Workflow for the project performed in the present work

4. Results and Discussion

4.1 Structure preparation

The first step was to download the six structures of CviR available on PDB. These structures were found with the help of the Biofilm Structural Database[79]. Information on each structure is available in the following table.

Table 2 - Available structures of CviR on PDB.

PDB Code	Protein	Resolution	Ligand	Strain
3QP1	Ligand-Binding Domain	1.55 Å	C6-HSL	Strain 31532
3QP2	Ligand-Binding Domain	1.638 Å	C8-HSL	Strain 31532
3QP4	Ligand-Binding Domain	1.55 Å	C10-HSL	Strain 31532
3QP5	Full Protein	3.249 Å	CL	Strain 31532
3QP6	Full Protein	2 Å	C6-HSL	Strain 12472
3QP8	Ligand-Binding Domain	1.6 Å	C10-HSL	Strain 12472

Different ligands with different activities are present in each structure. 3QP1 is complexed with its native ligand C6-HSL which is a full agonist in strain 31532. 3QP2 and 3QP4 are bonded to ligands with longer acyl chains which fail to fully activate CviR. C8-HSL leads to 40 % of the original activity and C10-HSL elicits only 6 %. These ligands also work as a partial antagonist in the presence of C6-HSL. 3QP5 is bonded to chlorolactone, an even stronger antagonist than C10-HSL. 3QP6 is bonded to C6-HSL which functions as an antagonist on strain 12472. Finally, 3QP8 is bonded to C10-HSL. This is a partial agonist, located closer to this strain's native ligand that is 3-hydroxy-C10-HSL[67].

All structures were opened in PyMOL. They were aligned using this program, and a monomer of each protein was isolated. After the alignment, the RMSD was calculated. The results are presented in table 3 and the alignment in figure 9. The analysis of this figure suggests that all structures are very similar, which is supported by the fact that all RMSD values are less than 1 Å.

Table 3 - RMSD values (Å) for all structures of CviR in PDB.

	3QP1	3QP2	3QP4	3QP5	3QP6	3QP8
3QP1	0.000	0.083	0.102	0.638	0.725	0.461
3QP2	0.083	0.000	0.107	0.578	0.635	0.487
3QP4	0.102	0.107	0.000	0.587	0.635	0.417
3QP5	0.638	0.578	0.587	0.000	0.939	0.765
3QP6	0.725	0.635	0.635	0.939	0.000	0.543
3QP8	0.461	0.417	0.417	0.765	0.543	0.000

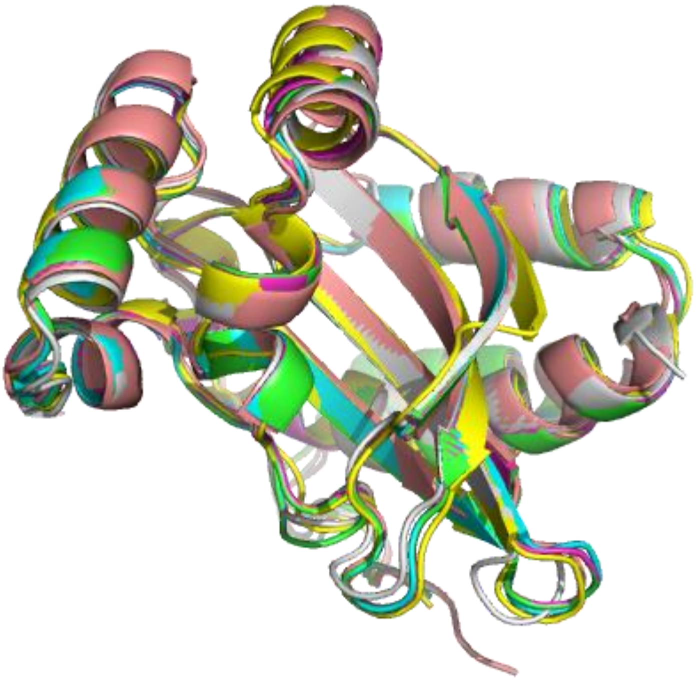


Figure 9- Alignment of all CviR structures, using the Cartoon representation, obtained using PyMOL.

The most important amino acids in the binding pocket are Tyr80, Trp84, Asp97 and Ser155. These amino acids also exhibited high similarity as seen on table 4 and figure 10. The only difference is with Tyr84 on 3QP5, which has a different conformation in relation with the other structures. The specific interactions between the protein, in this case 3QP1, and its native ligand are displayed in figure 11. This interactions map was obtained from the Biofilms Structural Database[79]. There are four hydrogen bonds connecting the ligand to the protein, one from each of the amino acids mentioned above.

Table 4 - RMSD values (Å) for the binding pockets of the available CviR structures.

	3QP1	3QP2	3QP4	3QP5	3QP6	3QP8
3QP1	0.000	0.053	0.070	0.770	0.770	0.768
3QP2	0.053	0.000	0.066	0.771	0.769	0.765
3QP4	0.070	0.066	0.000	0.768	0.769	0.769
3QP5	0.770	0.771	0.768	0.000	0.875	0.893
3QP6	0.770	0.769	0.769	0.875	0.000	0.134
3QP8	0.768	0.765	0.769	0.893	0.134	0.000

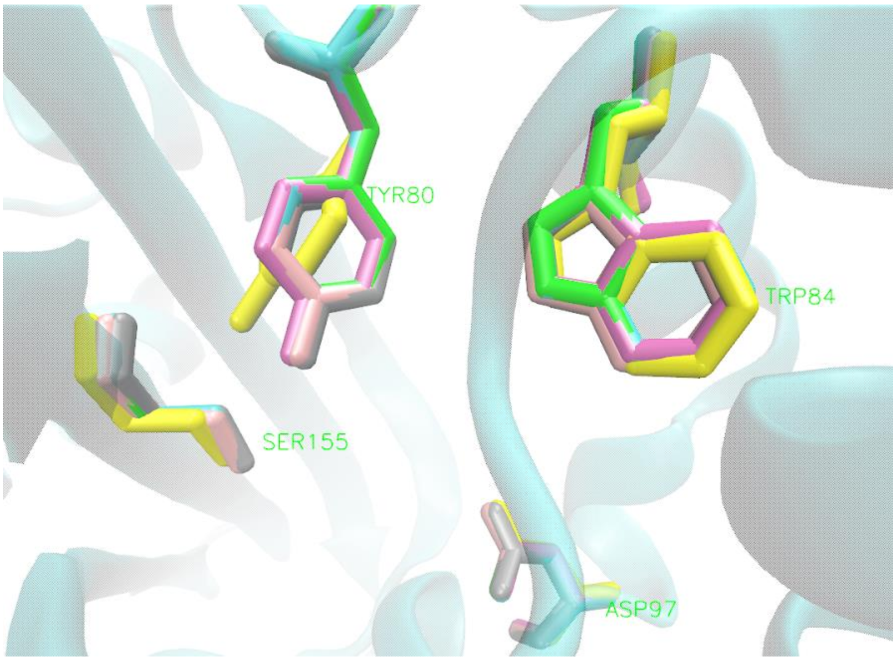


Figure 10 - Alignment of the binding pocket of all available CviR structures.

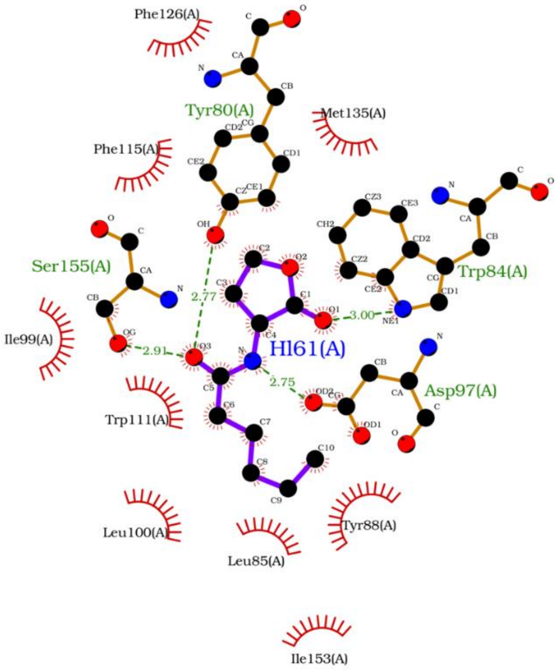


Figure 11 – Interactions map between 3QP1 and C6-HSL.

One important detail found in these structures is the importance of the amino acid resid 89 and its pose. On the CviR from strain 31532 (3QP1-3QP5) it is a methionine while on strain 12472 (3QP6 and 3QP8) it is a serine[67].

Starting with strain 31532, it was observed that Met89’s side chain swings away from the ligand-binding pocket along with an increase of length of ligand’s acyl chain. The extent of this variation relates to the antagonist capabilities of the ligand. 3QP1 is bonded to its native ligand and Met89 is in its original pose. In 3QP2, there are two conformations: one similar to 3QP1 and the other in which there is a small, intermediate variation on the side chain orientation. In both 3QP4 and 3QP5, the side chain has fully changed its position, swinging far away from the centre of the binding pocket. This is depicted on figure 12. This variation of the conformation of the side chain allows the bonding of larger ligands, which would not be possible otherwise[67].

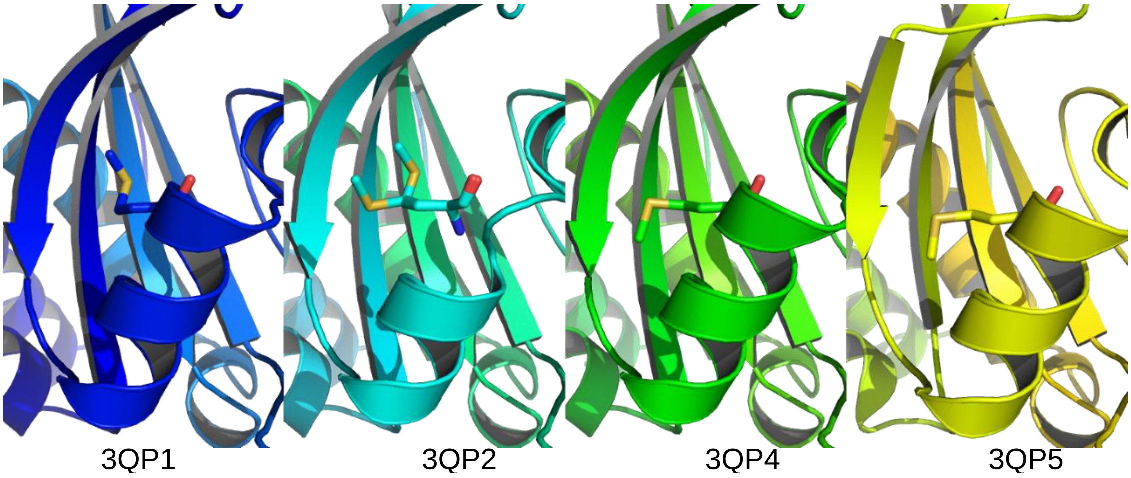


Figure 12- Variation of the Met89 side chain over the different structures.

On strain 12472 the presence of a serine, a smaller amino acid, means that larger ligands can be bonded to the protein without the need for a conformational change. This explains why the native ligand in this strain is a longer molecule. The bonding of smaller ligands originates a cavity in the binding pocket, which leads to a conformational change where the side chain of Met257 of the DNA binding domain is inserted. This interaction stabilizes a close conformation of the CviR dimer which explains the antagonist activity of smaller ligands like C6-HSL[67].

Before the molecular docking studies, using PyMOL, all water molecules were removed from each structure and the receptor and ligand were isolated in different *pdb* files. Finally, 3QP2 was separated into two different files, one for each conformation of Met89. 3QP2a has the intermediate conformation while 3QP2b has a similar conformation to 3QP1.

4.2 Redocking and Cross-Docking

4.2.1 Methods

4.2.1.1 Introduction

The first step towards selecting which structures and programs will be used in the virtual screening protocol is to perform redocking and cross-docking studies. The goal of the redocking studies is to see if the molecular docking programs are able to accurately reproduce the experimental complexed pose of the ligand. The RMSD between the re-docked and the original poses was calculated in order to better evaluate the results. In cross-docking studies, each ligand from each structure is docked into all other structures. Considering that all structures are of the same protein, this is an assessment of how accurately each structure docks the ligands from the other structures, i.e. its general usefulness.

To perform these studies, four programs were used, AutoDock 4, AutoDock Vina, LeDock and GOLD. The RMSD calculations were done using DockRMSD[161].

4.2.1.2 AutoDock 4

Each isolated ligand and receptor file was converted from *pdb* into *pdbqt* using two AutoDock 4 scripts, *prepare_ligand4.py* and *prepare_*receptor4*.py*, respectively.

In AutoDock 4, the energy of each probe atom is calculated and saved on grid maps, with each atom type in the ligand having one grid map. The grid parameter file sets the number of points in each dimension, the centre of the grid, the space between points, the types of probe atoms to be used, the filename of the receptor and the names of each grid map. To prepare this grid parameter file, the *prepare_gpf4.py* script was used. In order to generate the grid maps AutoGrid needs a user written input file containing the information on the centre of the grid, the size, and the spacing. Several grids were tested with the same centre but different sizes. The information on the best performing one is presented in table 5. After obtaining the grid parameter file and writing the input file, the grid maps can be generated using the script *autogrid4*.

Table 5 - Grid-box coordinates for AutoDock4.

AutoDock 4 Grid-box	
Centre x	19.790
Centre y	12.010
Centre z	51.400
Size x	40.000
Size y	40.000
Size z	37.300
Spacing	0.375

AutoDock 4 needs a specific docking parameter file for each ligand-receptor pair. This file was prepared using the *prepare_dpf4.py* script. For this work, the default Lamarckian genetic algorithm parameters were used.

After all the preparation, AutoDock 4 was run using the *autodock4* script. The result is a *dlg* file featuring all generated poses, for the ligand-receptor complex and their respective scores. In order to separate all structures into their own *pdbqt* files, the *write_all_complexes.py* script was used.

4.2.1.3 AutoDock Vina

The receptor and ligand *pdbqt* files were obtained in the same way than for AutoDock 4. Additionally, Vina also needs a docking parameters file. This file features information on the name of the receptor, the size, centre and spacing of the search space. To run Vina, the general Vina script is used. This script indicates which ligand should be used and the names of the two output files.

Several different runs were performed, using the same box centre, but different box sizes and different exhaustiveness values. Exhaustiveness relates to the time spent on each search. A higher value corresponds to more time spent. The default value is 8. Table 6 displays the best performing parameters.

Table 6 - Best performing AutoDock Vina parameters for the molecular system under study.

Vina Parameters	
Centre x	19.79
Centre y	12.01
Centre z	51.40
Size x	15.00
Size y	15.00
Size z	14.00
Spacing	1.00
Exhaustiveness	8.00

After running Vina, a log and out files are generated. The log file features information on the progression of the docking procedure, the final scores for each of the generated poses and the RMSD of every pose in reference to the best scored pose. The out file contains the structural information of all generated poses, in PDBQT format, with the highest scored pose in first place and all other poses following in decreasing order of the scoring.

4.2.1.4 GOLD

To perform the docking studies using GOLD, it is required a receptor file in *pdb* format and a ligand in mol2 format. Open Babel was used to convert the original ligand files from *pdb* to *mol2* format.

The first step was to add all hydrogens to the original *pdb* file. This originated a new protein structure in mol2 format, named *ID_protein.mol2* which is the one used for the docking procedure. The next step was to define the binding site. The centre coordinates used were the same selected in both AutoDock 4 and Vina. Several sphere radii were tested, with the final sphere volume being similar to all the different box sizes tested in both AutoDock 4 and Vina. The radii tested were 7 Å, 8 Å, 9 Å and 10 Å. After this, the ligand(s) to be docked were selected. The following step was to choose one from the four scoring functions available (ASP, CHEMPLP, CHEMSCORE or GOLDSCORE). All non-mentioned options were left on their default values.

Finally, GOLD can be run on the graphical interface or one can choose to run it on the background. By running it on the graphical interface, one can see the several log files being written as the docking procedure happens. On the other hand, by running on the background, there is no display of what is happening, but the docking files can be opened manually if necessary. A third option is to save all the parameters on a gold.conf file and running GOLD in the terminal using the previously mentioned *gold_auto* script. In this step of the work, GOLD was run using the graphical interface.

In the end, the best scored pose for each ligand, can be seen in the bestranking*.lst* file. Additionally, all the poses are written as *gold_soln_ligandname_m#_n.mol2* files*,* in which # is the number of the docking attempt. These are the most important output files by a GOLD run. However, there are others, such as the *rnk* files in which all poses for each ligand and their scores are displayed or *gold.err* files where every error is written (when one occurs).

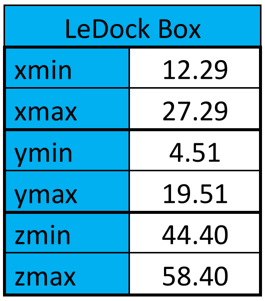
Table 7 - Optimized GOLD parameters for the molecular system under study.

GOLD Parameters	
Centre x	19.79
Centre y	12.01
Centre z	51.40
Sphere radius	9 Å

4.2.1.5 LeDock

LeDock requires the usage of the LePro application to prepare the receptor for the docking procedure. Using the *lepro_linux_x86* script, the original *pdb* file is transformed in a *pro.pdb.* All ligands have to be in *mol2* format, and so, all were converted from *pdb* to *mol2* using Open Babel. LeDock needs all the names of every ligand to be in a single text file, here named *list_ligand*. The final file required is the parameters file here named, *input_file.in.* This file features information on the name of the receptor file, the maximum RMSD desired between poses, the coordinates for the binding pocket, the number of desired binding poses and the name of the file containing the ligand's names. After preparing all input files, LeDock was run using the *ledock_linux_x86* script. Several box sizes were used, with the same sizes as in AutoDock 4 and Vina, with the coordinates of the best performing one displayed in table 8.

Table 8 – Optimized box coordinates for LeDock molecular docking calculations.



LeDock Box	
xmin	12.29
xmax	27.29
ymin	4.51
ymax	19.51
zmin	44.40
zmax	58.40

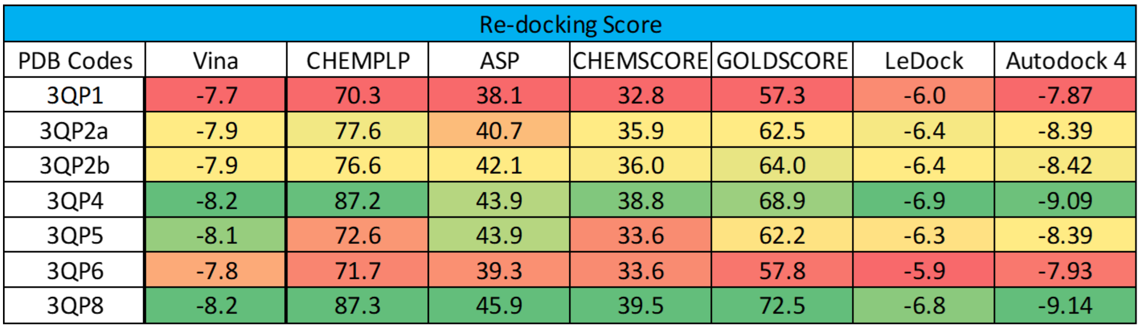
LeDock produces a *dok* output file, featuring the score for each generated pose, and their atomic coordinates. To open the generated poses in a molecular visualization software, the coordinates have to be converted into mol2 format using the *dok2mol2* script. To obtain the *docking_summary.txt,* which features the score of the best pose for each ligand, the script *ledock_anal* was used.

4.2.2 Results

4.2.2.1 Redocking

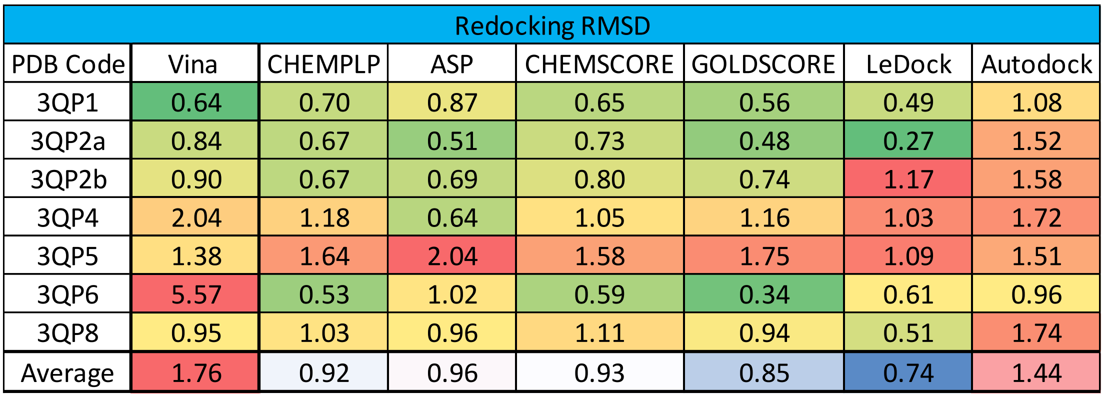
The scoring values for all the redocking studies are summarized in table 9. The corresponding RMSDs values are presented in table 10. For an easier interpretation of the results, a colour gradient is used in which the best results are coloured green and as the results get worse, the colour becomes closer to red. On the RMSD averages, the gradient is different, with the best results being coloured blue instead of green.

Table 9 - Redocking scores for all available CviR structures. Values for Vina and LeDock are in kcal/mol.



Re-docking Score							
PDB Codes	Vina	CHEMPLP	ASP	CHEMSCORE	GOLDSCORE	LeDock	Autodock 4
3QP1	-7.7	70.3	38.1	32.8	57.3	-6.0	-7.87
3QP2a	-7.9	77.6	40.7	35.9	62.5	-6.4	-8.39
3QP2b	-7.9	76.6	42.1	36.0	64.0	-6.4	-8.42
3QP4	-8.2	87.2	43.9	38.8	68.9	-6.9	-9.09
3QP5	-8.1	72.6	43.9	33.6	62.2	-6.3	-8.39
3QP6	-7.8	71.7	39.3	33.6	57.8	-5.9	-7.93
3QP8	-8.2	87.3	45.9	39.5	72.5	-6.8	-9.14

Table 10 - Redocking RMSD for all available CviR structures. All values are in Å



Redocking RMSD							
PDB Code	Vina	CHEMPLP	ASP	CHEMSCORE	GOLDSCORE	LeDock	Autodock
3QP1	0.64	0.70	0.87	0.65	0.56	0.49	1.08
3QP2a	0.84	0.67	0.51	0.73	0.48	0.27	1.52
3QP2b	0.90	0.67	0.69	0.80	0.74	1.17	1.58
3QP4	2.04	1.18	0.64	1.05	1.16	1.03	1.72
3QP5	1.38	1.64	2.04	1.58	1.75	1.09	1.51
3QP6	5.57	0.53	1.02	0.59	0.34	0.61	0.96
3QP8	0.95	1.03	0.96	1.11	0.94	0.51	1.74
Average	1.76	0.92	0.96	0.93	0.85	0.74	1.44

The worst scores are seen in 3QP1 and 3QP6, where the ligand is C6-HSL. This was expected because for strain 31532, C6-HSL is the native ligand while the other ligands are antagonists with higher affinity to CviR. The low score in 3QP6 can be explained with the substitution of Met89 by Ser89, which leads to a reduced affinity towards smaller molecules.

The RMSD shows that the best docking program, for reproducing experimental poses of the ligand is LeDock. On the other hand, the programs which performed worse were Vina and Autodock 4. However, the Vina's performance is not as bad as the average indicates. A better look at the results shows a reasonable value in all structures except in structure 3QP6 where Vina placed the molecule the opposite way, as seen on figure 12. The following images showing the superimposition of the re-docked and original poses, together with the RMSD values indicate a good overall performance from all the different molecular docking software. In general, AutoDock 4 proved to be less capable to accurately reproduce the experimental poses. This good performance is mainly seen with the lactone head group, with most programs struggling with the acyl chain.

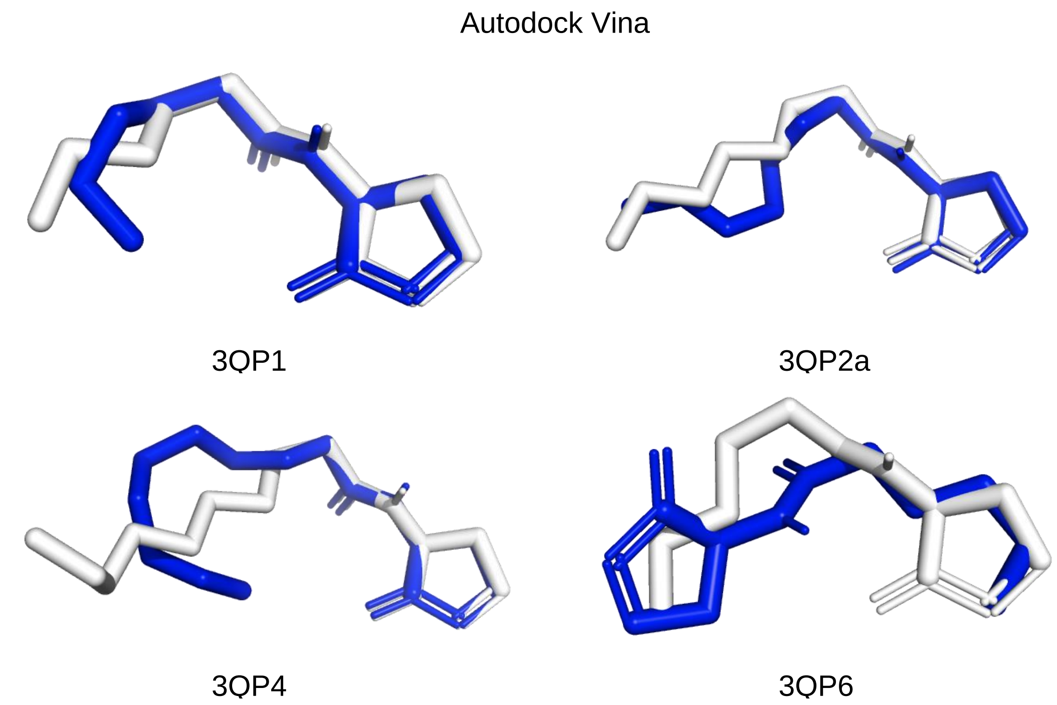


Figure 13 - Comparison between the original (white) and re-docked (blue) poses for several structures using AutoDock Vina.

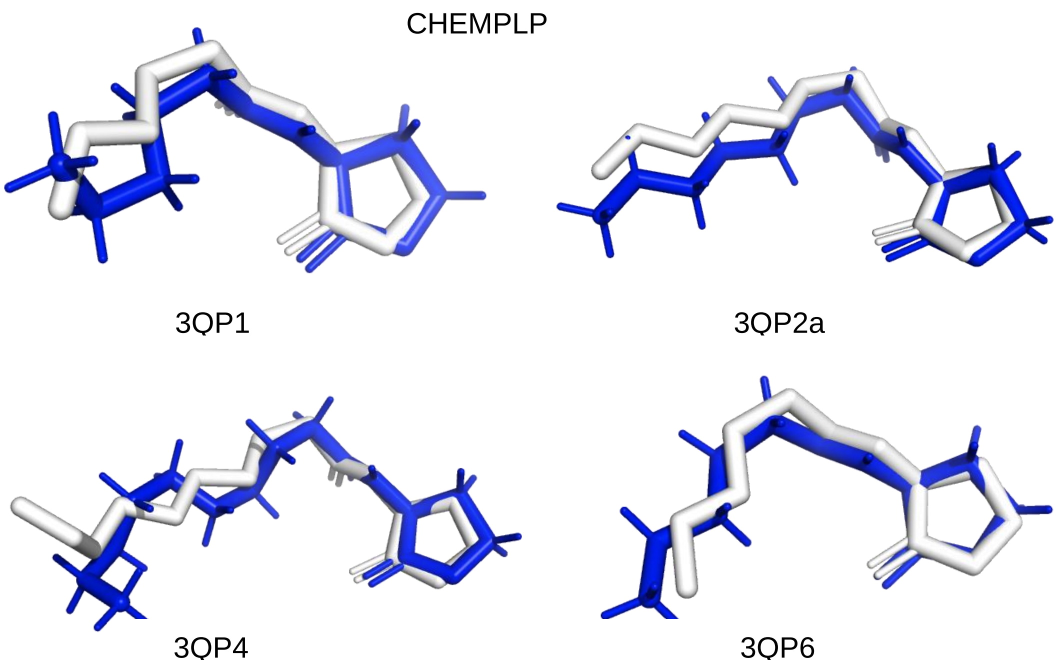


Figure 14 - Comparison between the original (white) and re-docked (blue) poses for several structures using CHEMPLP.

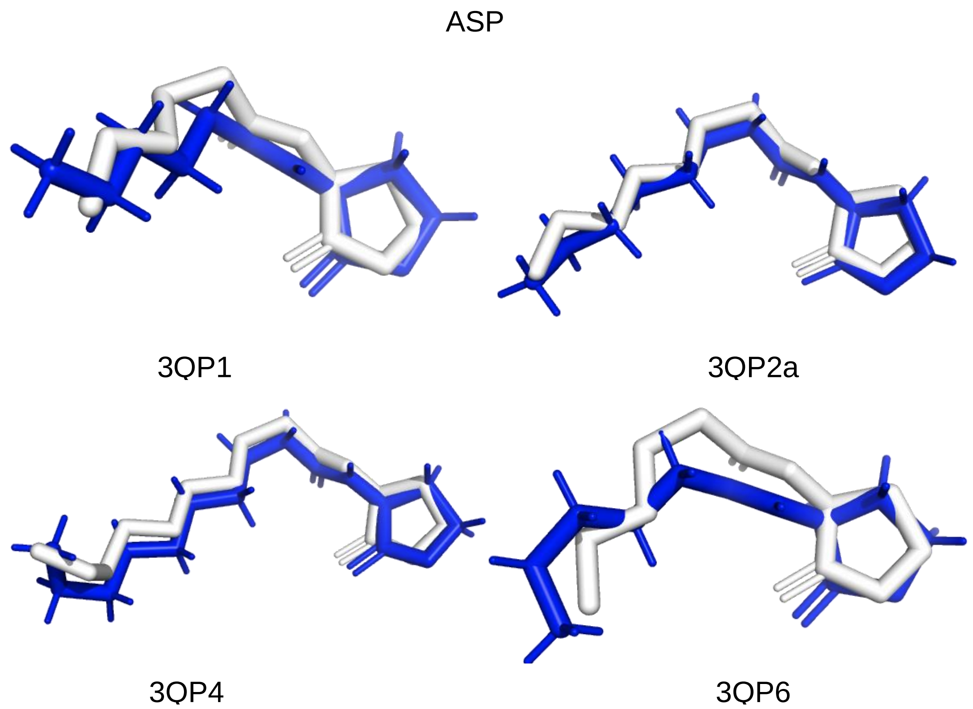


Figure 15 - Comparison between the original (white) and re-docked (blue) poses for several structures using ASP.

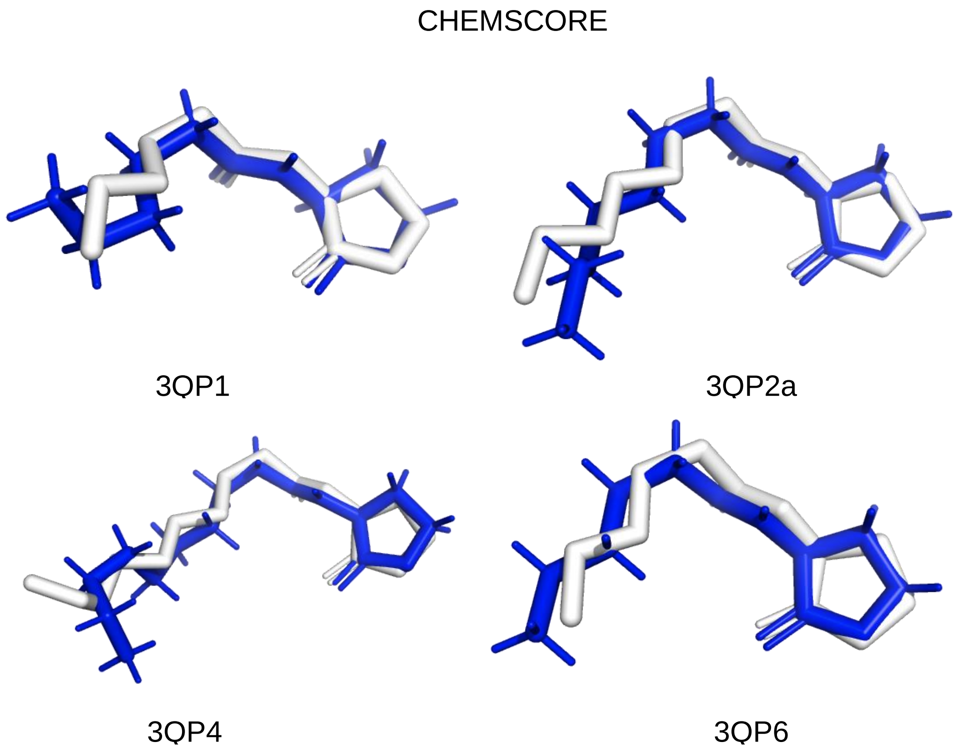


Figure 16 - Comparison between the original (white) and re-docked (blue) poses for several structures using CHEMSCORE.

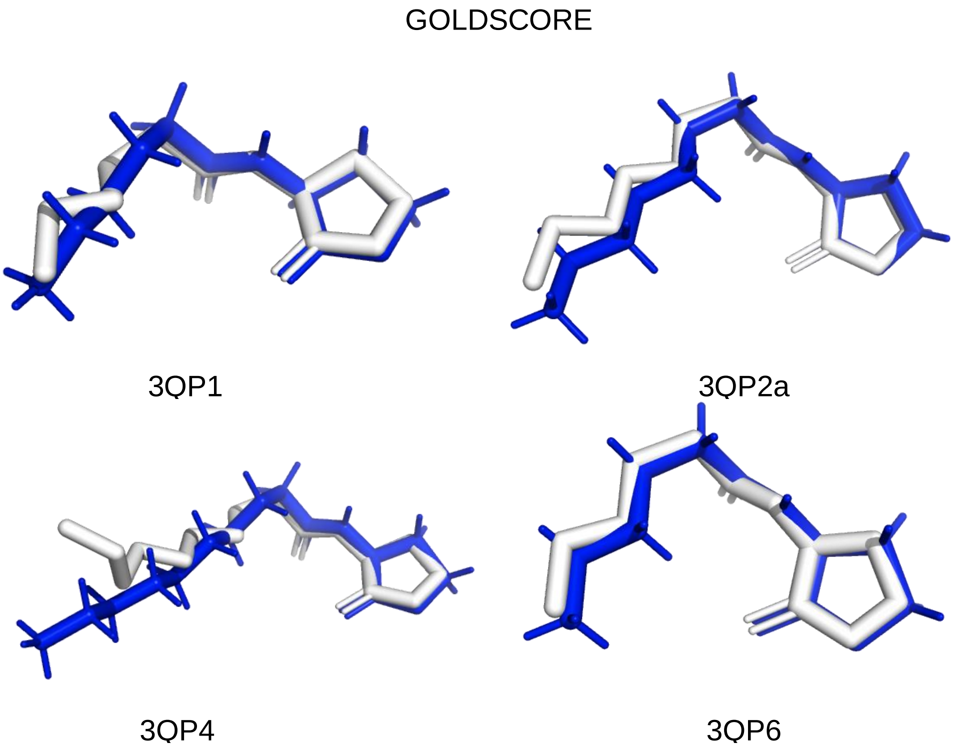


Figure 17 - Comparison between the original (white) and re-docked (blue) poses for several structures using GOLDSCORE.

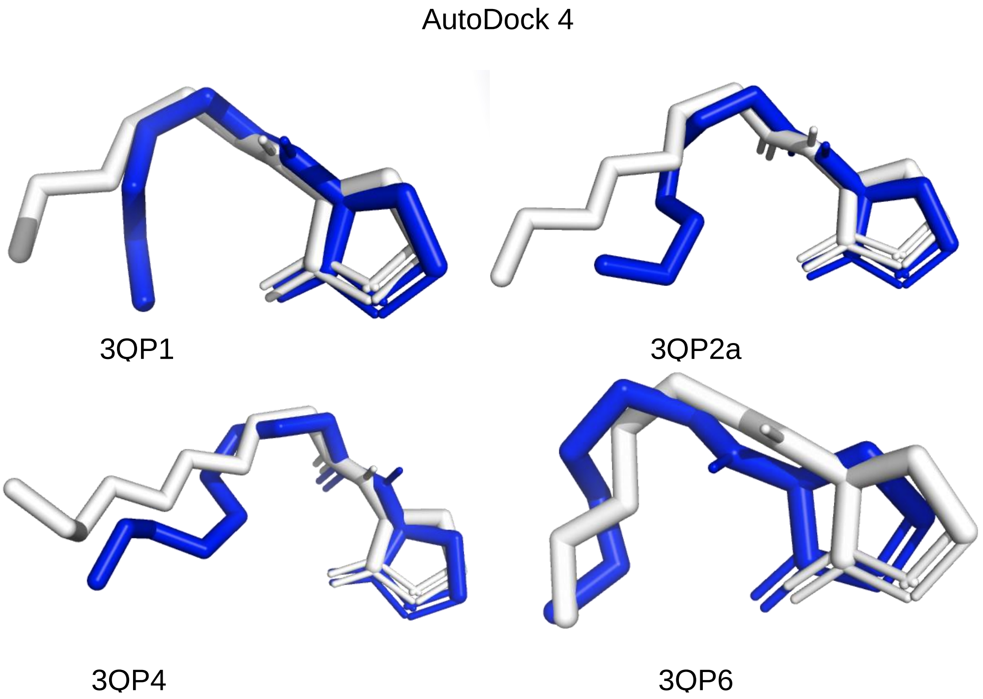


Figure 18 - Comparison between the original (white) and re-docked (blue) poses for several structures using AutoDock 4.

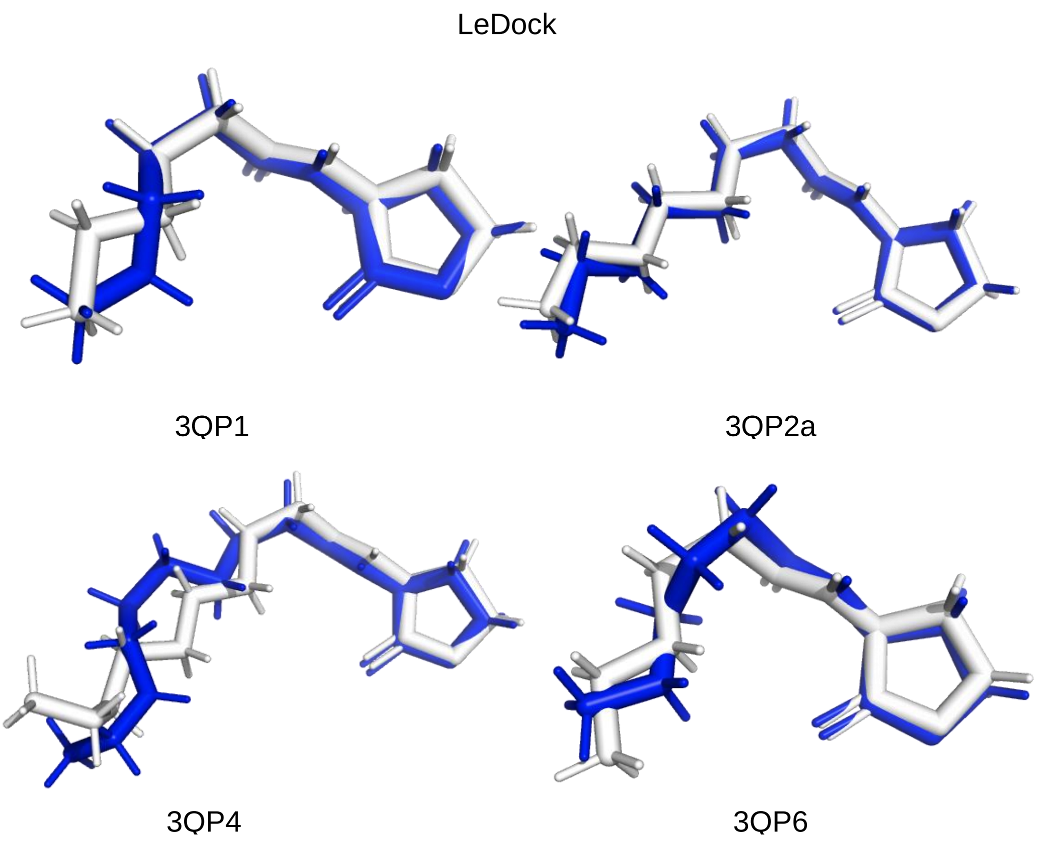


Figure 19 - Comparison between the original (white) and re-docked (blue) poses for several structures using LeDock.

4.2.2.2 Crossdocking

The Crossdocking scores for all software and structures are presented in the tables 11 to 17. As before, a colour gradient is applied a for better comprehension. The best scores are coloured green, and the worst are coloured in red, except for the averages where the best results are coloured in blue.

Table 11 - Crossdocking results for AutoDock Vina. All values are in kcal/mol.

AutoDock Vina								
PDB Code	Ligand	3QP1	3QP2a	3QP2b	3QP4	3QP5	3QP6	3QP8
3QP1	C6-HSL	-7.7	-7.6	-7.6	-7.8	-6.7	-7.7	-7.6
3QP2	C8-HSL	-8.0	-7.9	-7.9	-8.0	-7.3	-8.0	-7.9
3QP4	C10-HSL	-8.2	-8.2	-8.2	-8.2	-7.5	-8.3	-8.1
3QP5	CL	-8.4	-8.9	-8.9	-8.9	-8.1	-8.8	-9.0
3QP6	C6-HSL	-7.8	-7.7	-7.6	-7.7	-6.7	-7.7	-7.6
3QP8	C10-HSL	-8.2	-8.3	-8.2	-8.2	-7.4	-8.3	-8.2
	Average	-8.05	-8.10	-8.07	-8.13	-7.28	-8.13	-8.07

Table 12 - Crossdocking results for CHEMPLP.

CHEMPLP								
PDB Code	Ligand	3QP1	3QP2a	3QP2b	3QP4	3QP5	3QP6	3QP8
3QP1	C6-HSL	69.3	69.2	68.9	70.3	58.8	71.2	70.5
3QP2	C8-HSL	77.5	77.2	77.9	76.7	65.0	80.6	78.4
3QP4	C10-HSL	83.9	84.1	85.1	85.5	73.5	88.7	85.6
3QP5	CL	83.7	85.5	86.0	84.5	70.9	80.5	88.2
3QP6	C6-HSL	69.0	69.3	69.4	70.4	58.7	71.2	70.1
3QP8	C10-HSL	82.6	84.0	85.5	86.2	71.5	88.9	86.2
	Average	77.7	78.2	78.8	78.9	66.4	80.2	79.8

Table 13 - Crossdocking results for ASP.

ASP								
PDB Code	Ligand	3QP1	3QP2a	3QP2b	3QP4	3QP5	3QP6	3QP8
3QP1	C6-HSL	38.2	37.8	37.1	37.5	35.1	38.9	39.5
3QP2	C8-HSL	41.7	40.9	41.9	41.7	37.7	43.8	43.2
3QP4	C10-HSL	44.5	43.9	43.7	44.8	40.5	46.3	45.6
3QP5	CL	48.2	49.3	48.3	48.6	44.3	49.5	51.0
3QP6	C6-HSL	37.6	37.1	37.6	37.6	33.7	40.2	39.6
3QP8	C10-HSL	43.8	43.4	44.0	44.4	39.6	46.5	45.5
	Average	42.3	42.1	42.1	42.4	38.5	44.2	44.1

Table 14 - Crossdocking results for CHEMSCORE.

CHEMSCORE								
PDB Code	Ligand	3QP1	3QP2a	3QP2b	3QP4	3QP5	3QP6	3QP8
3QP1	C6-HSL	32.9	32.3	32.7	32.5	27.4	33.7	33.4
3QP2	C8-HSL	36.5	35.9	35.9	35.9	29.5	37.0	36.6
3QP4	C10-HSL	39.7	38.8	38.6	39.3	33.1	39.9	39.4
3QP5	CL	41.2	41.2	41.1	41.0	33.1	42.3	41.1
3QP6	C6-HSL	33.1	32.7	32.3	32.4	27.0	33.6	33.5
3QP8	C10-HSL	39.2	38.7	38.7	39.0	33.9	40.3	39.4
	Average	37.1	36.6	36.5	36.7	30.7	37.8	37.2

Table 15 - Crossdocking results for GOLDSCORE.

GOLDSCORE								
PDB Code	Ligand	3QP1	3QP2a	3QP2b	3QP4	3QP5	3QP6	3QP8
3QP1	C6-HSL	57.7	56.9	56.6	58.7	50.1	58.7	59.1
3QP2	C8-HSL	65.3	61.9	62.1	63.9	51.8	64.2	64.6
3QP4	C10-HSL	63.8	66.9	69.0	70.8	59.1	74.1	73.0
3QP5	CL	67.0	68.8	69.9	68.5	62.3	74.0	72.2
3QP6	C6-HSL	56.3	55.8	56.3	56.6	51.3	57.7	57.8
3QP8	C10-HSL	64.0	66.8	68.3	71.0	58.3	72.3	73.7
	Average	62.3	62.8	63.7	64.9	55.5	66.8	66.7

Table 16 - Crossdocking results for AutoDock 4 all values are in kcal/mol.

AutoDock 4								
PDB Code	Ligand	3QP1	3QP2a	3QP2b	3QP4	3QP5	3QP6	3QP8
3QP1	C6-HSL	-7.9	-7.8	-8.8	-7.9	-6.8	-7.9	-7.8
3QP2	C8-HSL	-8.4	-8.4	-8.5	-8.5	-7.5	-8.8	-8.6
3QP4	C10-HSL	-9.0	-8.9	-9.1	-9.1	-8.0	-9.2	-8.9
3QP5	CL	-9.1	-9.5	-9.4	-9.4	-8.4	-9.9	-9.7
3QP6	C6-HSL	-7.9	-7.8	-7.8	-7.9	-7.2	-7.9	-7.9
3QP8	C10-HSL	-8.9	-8.9	-9.0	-9.0	-7.9	-9.3	-9.0
	Average	-8.54	-8.53	-8.76	-8.63	-7.64	-8.84	-8.66

Table 17 - Crossdocking results for LeDock all values are in kcal/mol.

LeDock								
PDB Code	Ligand	3QP1	3QP2a	3QP2b	3QP4	3QP5	3QP6	3QP8
3QP1	C6-HSL	-5.99	-5.92	-5.91	-6.04	-5.02	-5.81	-5.78
3QP2	C8-HSL	-6.47	-6.41	-6.41	-6.53	-5.54	-6.30	-6.30
3QP4	C10-HSL	-6.36	-6.78	-6.67	-6.95	-5.85	-6.81	-6.68
3QP5	CL	-6.40	-7.32	-7.21	-7.43	-6.32	-7.26	-7.52
3QP6	C6-HSL	-6.01	-5.93	-5.92	-6.08	-5.09	-5.92	-5.81
3QP8	C10-HSL	-6.78	-6.82	-6.73	-7.00	-5.83	-6.94	-6.82
	Average	-6.34	-6.53	-6.48	-6.67	-5.61	-6.51	-6.49

The crossdocking studies show that, on average, the best results are seen on structures 3QP6 and 3QP8, both from strain 12472. From strain 31532 the structure which generates the best scores is 3QP4. In contrast, the worst results in all ligands and software are seen on structure 3QP5. This may be due to the bad resolution of this structure, which is by a considerable margin, the structure with the worse resolution, as seen on table 2.

Similar to the redocking studies, C6-HSL shows less affinity when compared to all other ligands, while C10-HSL and CL show the best results. This behaviour can be seen on the majority of structures and software.

One major difference from the redocking and crossdocking studies is the behaviour of 3QP6. As expected, the bad performance during the redocking studies was due to the ligand, C6-HSL. When docking other ligands, 3QP6 is frequently the best performing structure.

4.2.3 Conclusions

The redocking and crossdocking studies suggest that most software perform similarly with the worst being AutoDock 4 and, to a lesser extent, Vina. However, Vina is the easiest and fastest of all software used, while AutoDock 4 is the most time-consuming one. Because of this, the decision was made to not continue to use AutoDock 4 on the following steps of this work.

Regarding the structures, the main conclusions emerging from these results are:

- Both structures from strain 12472 (3QP6 and 3QP8) displayed the higher scores.
- 3QP6 is the structure that generated the best results.
- From strain 31532, 3QP4 displayed the higher and most consistent scores for all software.
- The structures 3QP1, 3QP2a and 3QP2b presented more variable scores.
- 3QP5, was consistently the worst performing of all available structures of CviR.

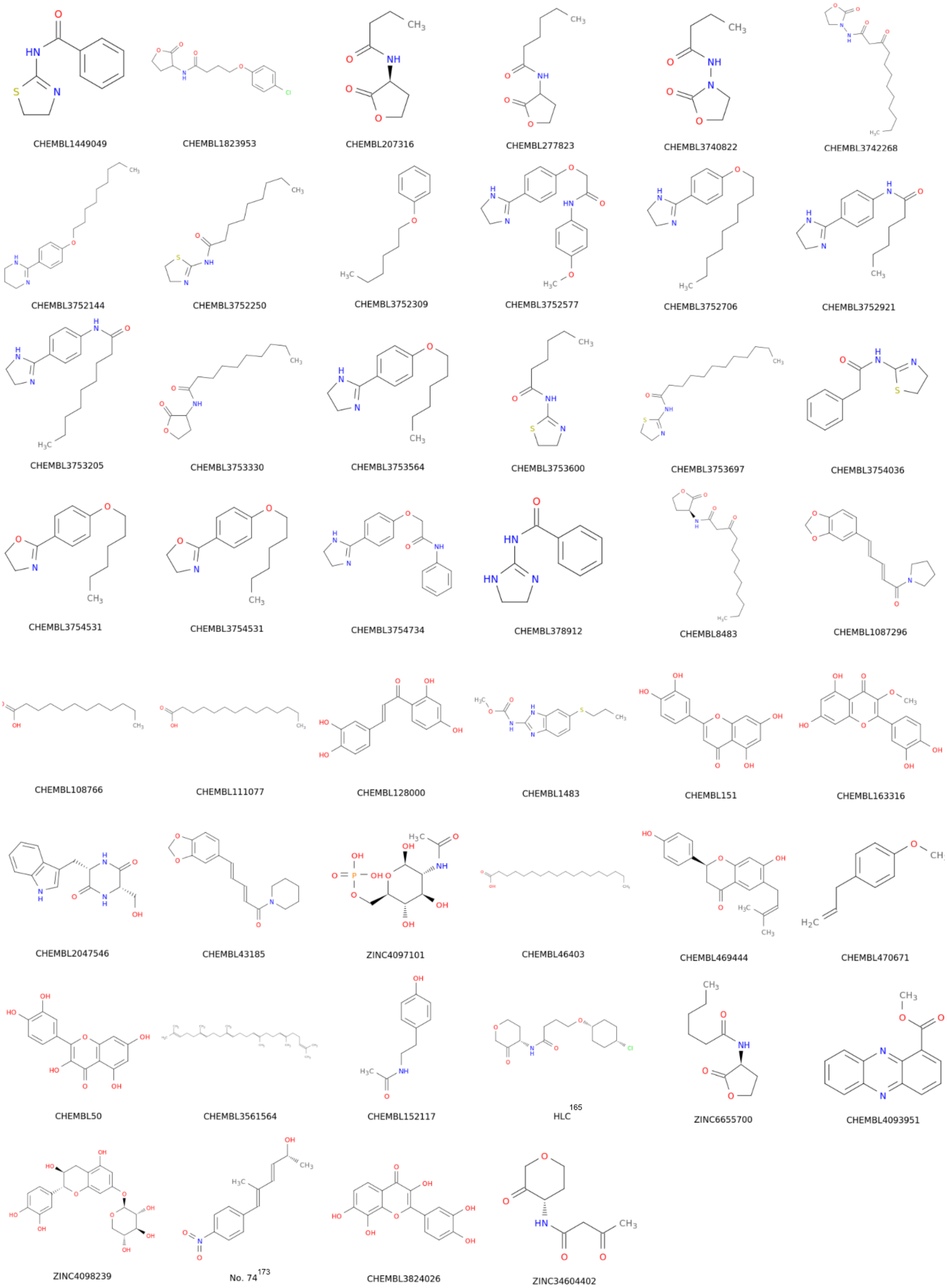
4.3 Optimization of the Virtual Screening Protocol

4.3.1 Methods

The optimization of the virtual screening protocol was separated into two parts. The first was a screening of known active compounds and the second a screening of actives and decoys.

The screening of active molecules was done on all previously mentioned molecular docking software, with the exception of AutoDock 4. Different sized binding pockets were used in this study, similarly to procedure adopted for the redocking and crossdocking studies. All software was run in the same way as in the previous studies. The only exception was GOLD, which from this point forward was run using the *gold_auto* script.

A total of 46 actives were used, 23 obtained from ChEMBL[141] and 23 were found in the literature[162–173] and downloaded from the PubChem[174]. The 46 actives can be seen on figure 20. Since all these compounds were downloaded in *smi* format, they were converted into *pdbqt* and *mol2* using Open Babel.



CHEMBL1449049 CHEMBL1823953 CHEMBL207316 CHEMBL277823 CHEMBL3740822 CHEMBL3742268

CHEMBL3752144 CHEMBL3752250 CHEMBL3752309 CHEMBL3752577 CHEMBL3752706 CHEMBL3752921

CHEMBL3753205 CHEMBL3753330 CHEMBL3753564 CHEMBL3753600 CHEMBL3753697 CHEMBL3754036

CHEMBL3754531 CHEMBL3754531 CHEMBL3754734 CHEMBL378912 CHEMBL8483 CHEMBL1087296

CHEMBL108766 CHEMBL111077 CHEMBL128000 CHEMBL1483 CHEMBL151 CHEMBL163316

CHEMBL2047546 CHEMBL43185 ZINC4097101 CHEMBL46403 CHEMBL469444 CHEMBL470671

CHEMBL50 CHEMBL3561564 CHEMBL152117 HLC[165] ZINC6655700 CHEMBL4093951

ZINC4098239 No. 74[173] CHEMBL3824026 ZINC34604402

Figure 20 – Known active molecules for CviR

The screening of actives and decoys was done only for the best performing binding pocket. To obtain the decoys, the 46 actives were submitted to the Database of Useful Decoys – Enhanced which generated 50 decoys per active. This led to a total of 2346 molecules.

The calculation of the metrics to evaluate the performance of VS protocols was performed using two different applications. The ROC and Enrichment Factor was calculated on Microsoft Excel. The web based application Screening Explorer was used for calculating Total Gain and BEDROC[143].

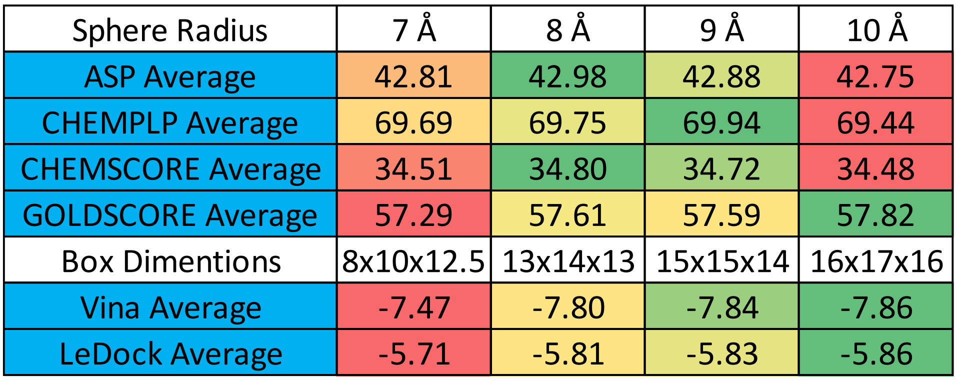
4.3.2 Results

4.3.2.1 Virtual Screening of the Active molecules

First all active compounds were docked on all structures, using all the binding pocket dimensions also used on the redocking studies. The average value of all the scores for each active on each structure was used to calculate an overall average score for each molecular docking program. These average scores are displayed on table 18.

The sphere used in GOLD and the box used in Vina and LeDock were designed to have similar volume. That way the performance of the different molecular docking software can be directly compared.

Table 18 - Average score of all actives on every structure. The scores for Vina and LeDock are expressed in kcal/mol. The box dimensions are expressed in Å.



Sphere Radius	7 Å	8 Å	9 Å	10 Å
ASP Average	42.81	42.98	42.88	42.75
CHEMPLP Average	69.69	69.75	69.94	69.44
CHEMSCORE Average	34.51	34.80	34.72	34.48
GOLDSCORE Average	57.29	57.61	57.59	57.82
Box Dimentions	8x10x12.5	13x14x13	15x15x14	16x17x16
Vina Average	-7.47	-7.80	-7.84	-7.86
LeDock Average	-5.71	-5.81	-5.83	-5.86

The results indicate that the overall best dimensions for the binding pocket chosen are a sphere radius of 9 Å for GOLD and a similar volume box for AutoDock Vina and LeDock.

The average values for each molecule (for each structure), using every molecular docking program were ranked. Figure 21 shows the average 10 best placed actives on the virtual screenings with different molecular docking programs.

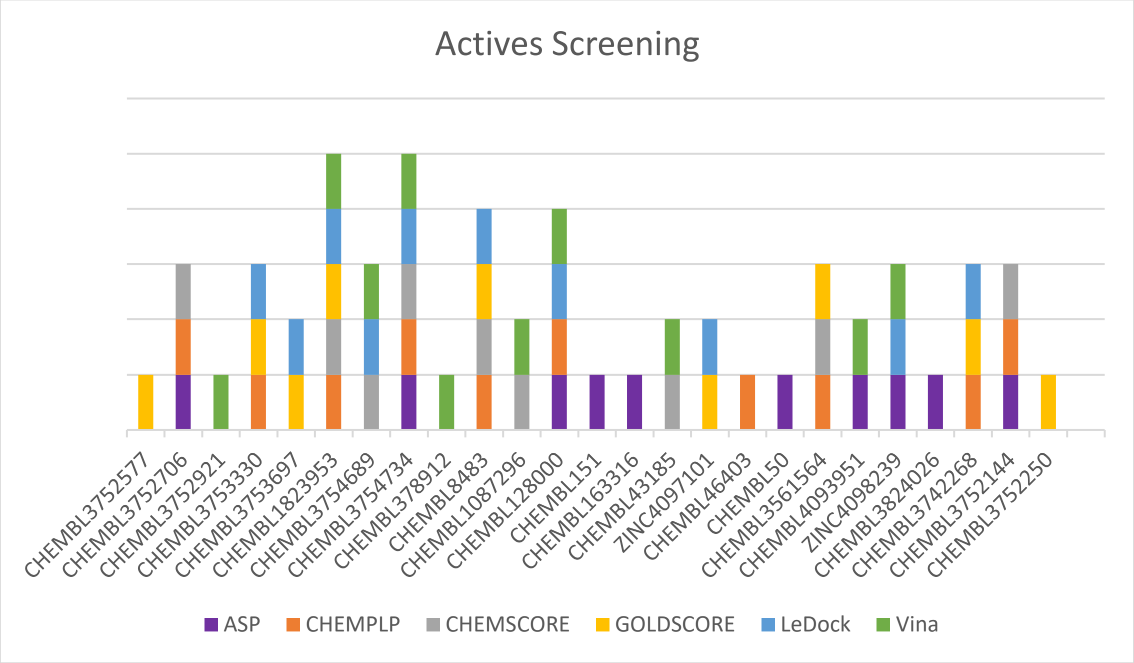


Figure 21 - Comparison of the 10 best placed active molecules on each virtual screening with different molecular docking programs.

Of the 46 active molecules in this virtual screening, 25 were placed at least once in the top 10 by one of the molecular docking programs. Of these 25 molecules, nine only appeared in the top 10 once, five appeared twice, seven appeared three times, two molecules appeared four and other two five times. Even though there was no molecule being placed in the top 10 in all programs, these results mean that there was some consistency between all scoring functions used.

Lastly, it was evaluated if some of the scoring functions have any bias towards ranking molecules with certain properties higher. For this purpose, the linear correlations between the scores and multiple chemical properties were calculated. The values for these properties were obtained and the statistical calculation was done using the open-source program for data analysis Datawarrior[175]. The results are presented on table 19.

Table 19 - Correlation between the used scoring functions and several chemical properties of the active molecules.

Bravais-Pearson Linear Correlation	Amides	Aromatic Nitrogens	Aromatic Rings	Basic Nitrogens
Score ASP	-0.21	0.12	0.62	0.29
Score CHEMPLP	-0.07	-0.20	-0.16	0.18
Score CHEMSCORE	0.00	0.10	-0.15	-0.27
Score GOLDSCORE	0.04	-0.07	-0.08	0.08
Score LeDock	-0.35	0.01	-0.04	-0.27
Score Vina	-0.08	-0.05	-0.58	-0.33
Bravais-Pearson Linear Correlation	Electronegative Atoms	H-Acceptors	H-Donors	Polar Surface Area
Score ASP	0.34	0.38	0.33	0.29
Score CHEMPLP	-0.16	-0.17	-0.24	-0.24
Score CHEMSCORE	0.22	0.24	0.32	0.34
Score GOLDSCORE	0.23	0.20	0.12	0.18
Score LeDock	-0.65	-0.58	-0.42	-0.56
Score Vina	-0.22	-0.24	-0.12	-0.11
Bravais-Pearson Linear Correlation	Rotatable Bonds	Total Molweight	Total Surface Area	cLogP
Score ASP	-0.05	0.69	0.52	0.18
Score CHEMPLP	0.66	0.58	0.77	0.58
Score CHEMSCORE	-0.30	-0.46	-0.58	-0.50
Score GOLDSCORE	0.56	0.74	0.78	0.33
Score LeDock	-0.08	-0.46	-0.28	0.32
Score Vina	0.30	-0.32	-0.17	0.02

Before analysing the results, it is important to remember than Vina and LeDock have scores with negative values while GOLD has scores with positive values. The ASP scoring function generates better results for molecules with a higher number of aromatic rings, a higher molecular mass and total surface area. CHEMPLP also displays a bias towards molecules with a higher molecular mass and surface area, but also favours molecules with a higher number of rotatable bonds. On the other hand, CHEMSCORE displays better results for molecules with a lower molecular mass, surface area and partition coefficient. GOLDSCORE shows favouritism towards molecules with higher molecular mass, total surface area, and number of rotatable bonds. All GOLD scoring functions display a bias towards molecules with a higher molecular mass when compared to LeDock and Vina. LeDock's scoring function ranks favourably molecules with a higher number of electronegative atoms, hydrogen bond acceptors and polar surface area. Finally, Vina shows less bias on relation to the different properties when compared with the other scoring functions, with the only substantial correlation being with a higher number of aromatic rings.

4.3.2.2 Actives vs Decoys Virtual Screening

The 2346 molecules were screened on all available structures using Vina, GOLD and LeDock. The results were evaluated using several metrics. The results obtained for the area under the ROC curve (ROC) metrics, are presented in table 20.

Table 20 - Area under the curve from the actives vs decoys virtual screening.

AUC							
PDB Code	Vina	CHEMPLP	ASP	CHEMSCORE	GOLDSCORE	LeDock	Average
3QP1	0.746	0.703	0.544	0.676	0.668	0.701	0.673
3QP2a	0.739	0.687	0.551	0.682	0.650	0.710	0.670
3QP2b	0.752	0.698	0.542	0.677	0.647	0.718	0.672
3QP4	0.772	0.712	0.580	0.696	0.672	0.744	0.696
3QP5	0.619	0.567	0.490	0.549	0.503	0.565	0.549
3QP6	0.820	0.757	0.614	0.715	0.722	0.753	0.730
3QP8	0.817	0.762	0.609	0.710	0.708	0.743	0.725
Average	0.752	0.698	0.561	0.672	0.653	0.705	

These results show that the molecular docking program capable of better discriminating the active molecules from the decoys is AutoDock Vina. This program presented, not only the higher average AUC, but also the higher AUC across every structure. Other programs displaying a good performance are GOLD, using the CHEMPLP scoring function, and LeDock. On the other side of the scale, ASP is, by a clearly, the scoring function with the lowest AUC, generating the lowest value in every structure. The average AUC for ASP is 0.092 lower than the next lowest value, which is GOLDSCORE.

In the same way as the crossdocking studies, the structures that led to better AUC results were 3QP6 and 3QP8. 3QP6 was the overall best performing structure, having the best results in every program. There was only one exception, where 3QP8 presented the higher value of AUC. Between the structures from strain 31532, 3QP4 generated the highest values in every program, with 3QP1, 3QP2a and 3QP2b displaying similar performance. In contrast, 3QP5 consistently exhibited the worst AUC values, with its average AUC being 0.121 lower than 3QP2a that presented the next lowest values.

An overall view of the ROC curves for the active vs decoys virtual screening using 3QP6 can be seen on figure 22.

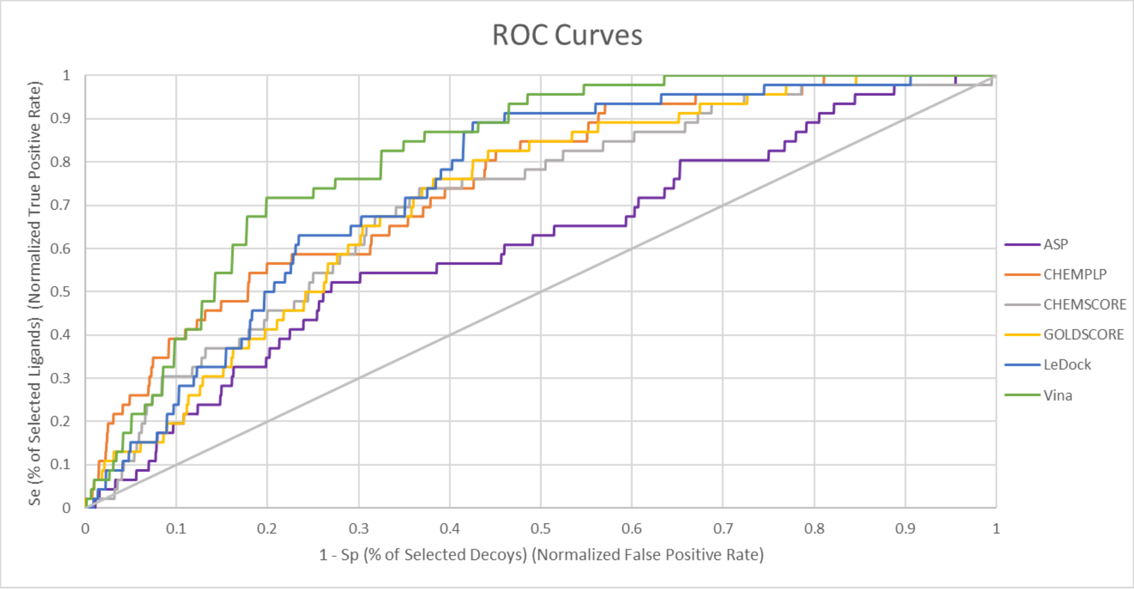


Figure 22 - ROC curve of the active vs decoys virtual screening for 3QP6.

The results, obtained for the enrichment factor (EF) at 1 %, are presented in table 21.

Table 21 - Enrichment factor at 1 % of the actives vs decoys virtual screening.

Receptor	1%											
	Vina		CHEMPLP		ASP		CHEMSCORE		GOLDSCORE		LeDock	
	EF	nº Actives/46	EF	nº Actives/46	EF	nº Actives/46	Score	nº Actives/46	EF	nº Actives/46	EF	nº Actives/46
3QP1	2.22	1	11.08	5	0.00	0	2.22	1	0.00	0	0.00	0
3QP2a	2.22	1	2.22	1	0.00	0	2.22	1	2.22	1	0.00	0
3QP2b	3.92	2	4.43	2	0.00	0	2.22	1	2.22	1	0.00	0
3QP4	2.22	1	4.43	2	0.00	0	2.22	1	6.65	3	0.00	0
3QP5	0.00	0	2.22	1	0.00	0	2.22	1	0.00	0	0.00	0
3QP6	6.65	3	4.43	3	0.00	0	2.22	1	4.43	3	2.22	1
3QP8	4.43	2	4.43	2	2.22	1	2.22	1	6.65	3	0.00	0
Average	3.09	1.43	4.75	2.29	0.32	0.14	2.22	1.00	3.17	1.57	0.32	0.14

The enrichment factor at 1% evaluates how many active molecules each docking procedure places within the top 1% of the results. In this database, 1% corresponds to the 23 best scored molecules.

The best performing program according to this metric is GOLD using the CHEMPLP scoring function, the only program capable of placing a least on molecule on the top 1 % (corresponding to a EF value of 2.22) in every structure, and resulting on the highest obtained EF value at 1 % 11.08 (corresponding to 5 molecules). CHEMPLP was followed by the GOLDSCORE and AutoDock Vina. The worst performing programs are ASP and LeDock. Both programs were only able to place one active molecule on the top 1 % in one of the structures.

Using the EF at 1 % metrics, the structures with that produced the best results were, as before, both structures from strain 12472. Additionally, 3QP6 presented a better average value. Concerning the other structures, 3QP1 and 3QP4 had the highest EF values, with 3QP1 using CHEMPLP displaying the highest overall score, placing 5 active molecules in the top 1 %.

The results, obtained for the enrichment factor at 5 % metric, are presented on table 22.

Table 22 - Enrichment factor at 5 % of the actives vs decoys virtual screening.

Receptor	5%											
	Vina		CHEMPLP		ASP		CHEMSCORE		GOLDSCORE		LeDock	
	EF	nº Actives/46	EF	nº Actives/46	EF	nº Actives/46	EF	nº Actives/46	EF	nº Actives/46	EF	nº Actives/46
3QP1	2.18	5	4.79	11	1.31	3	1.74	4	1.74	5	1.74	4.00
3QP2a	3.92	9	5.23	12	1.74	4	1.31	3	1.74	4	2.62	6.00
3QP2b	3.92	9	5.66	13	1.74	4	2.18	5	2.61	6	4.36	10.00
3QP4	4.36	10	4.79	11	1.31	3	2.18	5	2.61	6	4.36	10.00
3QP5	0.87	2	0.87	2	0.87	2	0.87	2	0.44	1	0.44	1.00
3QP6	3.49	8	5.23	12	1.31	3	2.18	5	2.61	6	3.05	7.00
3QP8	4.36	10	5.23	12	1.74	4	2.61	6	3.05	7	2.18	5.00
Average	3.30	7.57	4.54	10.43	1.43	3.3	1.87	4.29	2.12	5.00	2.68	6.14

The enrichment factor at 5 % evaluates how many active molecules each docking procedure places within the top 5 % of the results. In this database, the number of the 5 % best scored molecules was 117.

The best performing program was GOLD using the CHEMPLP scoring function, with a maximum of 13 active molecules found, using 3QP2b. After GOLD, the best results were obtained by AutoDock Vina and LeDock. As before, ASP generated the worst results, only being able to place 2 molecules on the top 5 % in 3QP5.

Unlike the previous results, the best performing structures was 3QP2b, followed by 3QP4, and then the two structures from strain 12472, led by 3QP8. The worst values were obtained, as it has become usual, using 3QP5.

The results, obtained for the enrichment factor (EF) at 20 % metrics, are presented in table 23.

Table 23 - Enrichment factor at 20 % of the actives vs decoys virtual screening.

Receptor	20%											
	Vina		CHEMPLP		ASP		CHEMSCORE		GOLDSCORE		LeDock	
	EF	nº Actives/46	EF	nº Actives/46	EF	nº Actives/46	EF	nº Actives/46	EF	nº Actives/46	EF	nº Actives/46
3QP1	2.60	24	2.06	19	1.19	11	1.95	18	1.41	13	2.39	22.00
3QP2a	2.28	21	2.28	21	1.30	12	2.17	20	1.95	18	2.17	20.00
3QP2b	2.50	23	2.28	21	1.30	12	2.28	21	1.52	14	2.29	22.00
3QP4	2.50	23	2.39	22	1.74	16	2.39	22	1.63	15	2.39	22.00
3QP5	1.19	11	1.30	12	1.19	11	1.08	10	0.76	7	1.09	10.00
3QP6	3.58	33	2.82	26	1.74	16	2.28	21	2.06	19	2.50	23.00
3QP8	3.04	28	2.82	26	1.84	17	2.06	19	2.49	23	2.39	22.00
Average	2.53	23.29	2.28	21.00	1.47	13.6	2.03	18.71	1.69	15.57	2.17	20.14

The enrichment factor at 20 % evaluates how many active molecules each docking procedure places within the top 20 % of the results. In this database, the number of the 20 % best scored molecules was 469.

The results indicate that the program which is more capable for placing active molecules on the top 20 % is AutoDock Vina, with a maximum of 33 active molecules being found on 3QP6. GOLD using CHEMPLP also had a good performance, and so did LeDock. On the other hand, ASP continues to be the worst performing molecular docking software in this work, in 3QP1 it was only being able to correctly place 11 molecules in the top 20%.

Concerning the structures, the highest EF 20 % values were obtained for 3QP6, closely followed by 3QP8. As for the other structures, 3QP4 continues to display the best results, while 3QP5 also continues to have the worst performance.

The results obtained for the Total gain metrics are presented in table 24.

Table 24 - Total Gain values for the actives vs decoys virtual screening.

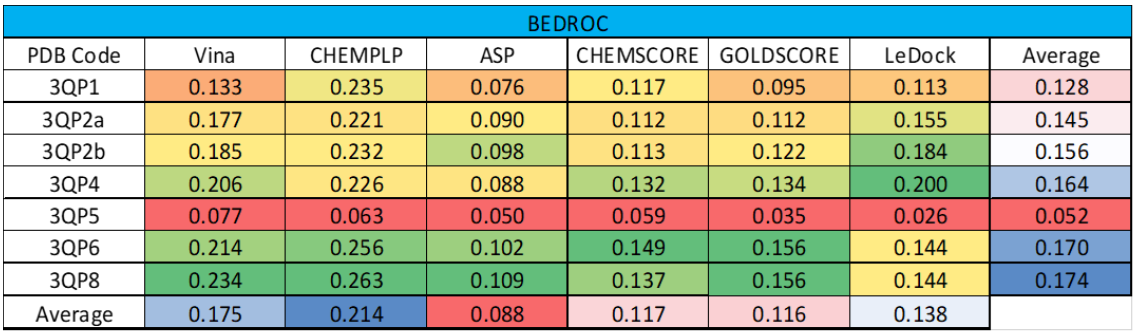
Total Gain							
PDB Code	Vina	CHEMPLP	ASP	CHEMSCORE	GOLDSCORE	LeDock	Average
3QP1	0.297	0.310	0.074	0.220	0.231	0.261	0.232
3QP2a	0.311	0.268	0.075	0.212	0.212	0.283	0.227
3QP2b	0.330	0.287	0.067	0.215	0.219	0.298	0.236
3QP4	0.358	0.300	0.113	0.230	0.249	0.335	0.264
3QP5	0.134	0.093	0.017	0.065	0.012	0.067	0.065
3QP6	0.420	0.350	0.159	0.260	0.304	0.312	0.301
3QP8	0.417	0.349	0.155	0.254	0.291	0.309	0.296
Average	0.324	0.280	0.094	0.208	0.217	0.266	

Similar to the results obtained from the other metrics, the best total gain values are obtained using AutoDock Vina, GOLD with the CHEMPLP scoring function, and LeDock. In this case, AutoDock Vina has the highest average total gain, having the absolute overall value in six of the seven structures. The worst values are once again obtained using GOLD with the ASP scoring function, with its total gain value being 0.114 inferior to the next lowest performing software.

The best performing structure was 3QP6, once again followed by 3QP8. Using AutoDock Vina with these two structures resulted in the only two cases in which the total gain value was higher than 0.4. Among the other structures, 3QP4 remains the structure displaying the best results. On the other hand, 3QP5 remains the worst performing structure, with a total gain value 0.162 inferior then the next worst score.

The results, obtained for BEDROC metrics, are presented in table 25.

Table 25 - BEDROC values for the actives vs decoys virtual screening.



BEDROC							
PDB Code	Vina	CHEMPLP	ASP	CHEMSCORE	GOLDSCORE	LeDock	Average
3QP1	0.133	0.235	0.076	0.117	0.095	0.113	0.128
3QP2a	0.177	0.221	0.090	0.112	0.112	0.155	0.145
3QP2b	0.185	0.232	0.098	0.113	0.122	0.184	0.156
3QP4	0.206	0.226	0.088	0.132	0.134	0.200	0.164
3QP5	0.077	0.063	0.050	0.059	0.035	0.026	0.052
3QP6	0.214	0.256	0.102	0.149	0.156	0.144	0.170
3QP8	0.234	0.263	0.109	0.137	0.156	0.144	0.174
Average	0.175	0.214	0.088	0.117	0.116	0.138	

The last metric used was BEDROC, a normalization of the robust initial enhancement, used to evaluate the early recognition of actives. The three best performing programs according to this metric were, as in most other metrics, GOLD with the CHEMPLP scoring function, AutoDock Vina and LeDock. In this case, CHEMPLP displayed the higher values. In agreement with the other metrics, the ASP scoring function once again generated the lowest values.

Finally, the structures which displayed the best BEDROC value were, as in most other metrics, both structures from strain 12472. In this case, the active vs decoys virtual screenings using 3QP8 generated the highest average value. 3QP4 was the best performing structure among the structures from strain 31532. 3QP5 continued to be the structure that led to the worst scores.

4.3.3 Conclusions

The objective of this section of the work was to develop an optimized virtual screening protocol that will be applied to larger databases. This way the probability to find promising molecules in further VS experiments was increased.

Three main components were evaluated through all metrics above displayed: the size of the binding area used for the molecular docking, the different molecular docking programs and all the available structures of CviR.

Concerning the size of the grid box, or sphere, depending on what software is used, the more consistent high scores were obtained for a radius of 9 Å in GOLD, and a box with the dimensions present in table 26 for AutoDock Vina and LeDock. Considering that these dimensions were generating the better scores, they were used for the Virtual Screening experiments.

Table 26 - Grid-box dimensions for AutoDock Vina and LeDock, values are expressed in Å.

Box dimensions	
Size x	15.0
Size y	15.0
Size z	14.0

The best performing programs were consistent across most of the metrics, with AutoDock Vina and GOLD using CHEMPLP being more capable of distinguishing the active compounds from the decoys, generally followed by LeDock. While LeDock usually generates lower results, principally with a very low EF at 1 %, it generates better results in all other metrics. Additionally, it and has the advantage that it generates results that are very easy to analyse. With this in mind, the software used in the Virtual Screening stage of this work were AutoDock Vina, GOLD with the CHEMPLP scoring function (GOLD/CHEMPLP) and LeDock.

Finally, it is important to select the receptor structures. Across nearly all metrics, the best results were obtained using 3QP6 and 3QP8. However, since they are both from strain 12472, the decision was made to use the best performing structure from each strain. The structure form strain 31532 that generally generated the best results was 3QP4. Therefore, the two CviR structures used in the Virtual Screening step of this work were 3QP4 and 3QP6.

4.4 Virtual Screening

4.4.1 Methods

The virtual screening protocol followed the conclusions obtained in the optimization step of this work. AutoDock Vina, GOLD/CHEMPLP and LeDock were run in the conditions optimized as in the previous chapter. These molecular docking procedures and were applied to two databases: a DrugBank FDA (U.S. Food and Drug

Administration) approved database available on ZINC, and Mu.Ta.Lig. Virtual Chemotheca.

ZINC is a public access database, which is used to obtain compounds for several uses. These include virtual screening, ligand discovery and force field development. It was created in the Department of Pharmaceutical Chemistry at the University of California (UCSF) and includes 220 million molecules[176]. In the present work, 1657 FDA approved molecules were downloaded in SMILES format from this database. These molecules were further converted into *pdbqt* and *mol2* using Open Babel*.* For simplicity, this database will be referred as ZINC/FDA.

The Mu.Ta.Lig. (Multi-Target Ligand) Chemotheca database was developed with the goal of identifying multi-target agents and repurposing known active compounds. For simplicity, this database will be referred just as Chemotheca. A large number of molecules with promising pharmaceutical relevance are developed every year and are forgotten when they fail to have the desired effect. However, these molecules can have a positive effect on other targets. The availability of a list of inactive compounds for a specific target can be useful for generating decoys or developing QSAR models. For this purpose, Chemotheca allows users to not only download compounds but also to upload their own[177]. At the time of this work, this database features 64804 compounds. These compounds were downloaded in *sdf* format and converted into *pdbqt* and *mol2* formats using Open Babel.

4.4.2 Results

4.4.2.1 ZINC/FDA database

Table 27 displays the 20 best ranked molecules for the virtual screening of the ZINC/FDA compounds database, using AutoDock Vina for both 3QP4 and 3QP6. Considering that all the molecules on this database are FDA approved drugs, they are identified by the name used in Drugbank. Histograms for these virtual screening simulations are present in figure 23.

Table 27 - Results of the virtual screening on the ZINC/FDA database using AutoDock Vina. The scores are expressed in kcal/mol.

Virtual Screening Results					
3QP4			3QP6		
Rank	Ligand	Score	Rank	Ligand	Score
1	Atovaquone	-12.1	1	Atovaquone	-11.9
2	Risperidone	-11.1	2	Mebendazole	-11.1
3	Mebendazole	-11.0	3	Risperidone	-10.9
4	Paliperidone	-10.5	4	Paliperidone	-10.9
5	Tolnaftate	-10.4	5	Benzoyl peroxide	-10.4
6	Axitinib	-10.4	6	Droperidol	-10.4
7	Umeclidinium	-10.4	7	Permethrin	-10.3
8	Lansoprazole	-10.3	8	Pimozide	-10.3
9	Permethrin	-10.2	9	Bicalutamide	-10.2
10	Belinostat	-10.2	10	Ketoprofen	-10.1
11	Benzoyl peroxide	-10.1	11	Nebivolol	-9.9
12	Fenofibric acid	-10.1	12	Cilostazol	-9.9
13	Ketoprofen	-10.0	13	Vismodegib	-9.9
14	Dexketoprofen	-10.0	14	Pirfenidone	-9.8
15	Haloperidol	-10.0	15	Nateglinide	-9.8
16	Fenofibrate	-10.0	16	Mirabegron	-9.8
17	Cilostazol	-10.0	17	Niraparib	-9.8
18	Estradiol	-10.0	18	Isocarboxazid	-9.7
19	Dolasetron	-9.9	19	Benzophenone	-9.7
20	Nateglinide	-9.9	20	Sulfasalazine	-9.7
21	Benzophenone	-9.9	21	Iloperidone	-9.7
22	Nebivolol	-9.9	22	Dapsone	-9.6
23	Iloperidone	-9.8	23	Tolnaftate	-9.6
24	Eletriptan	-9.8	24	Haloperidol	-9.6
25	Vemurafenib	-9.8	25	Ziprasidone	-9.6

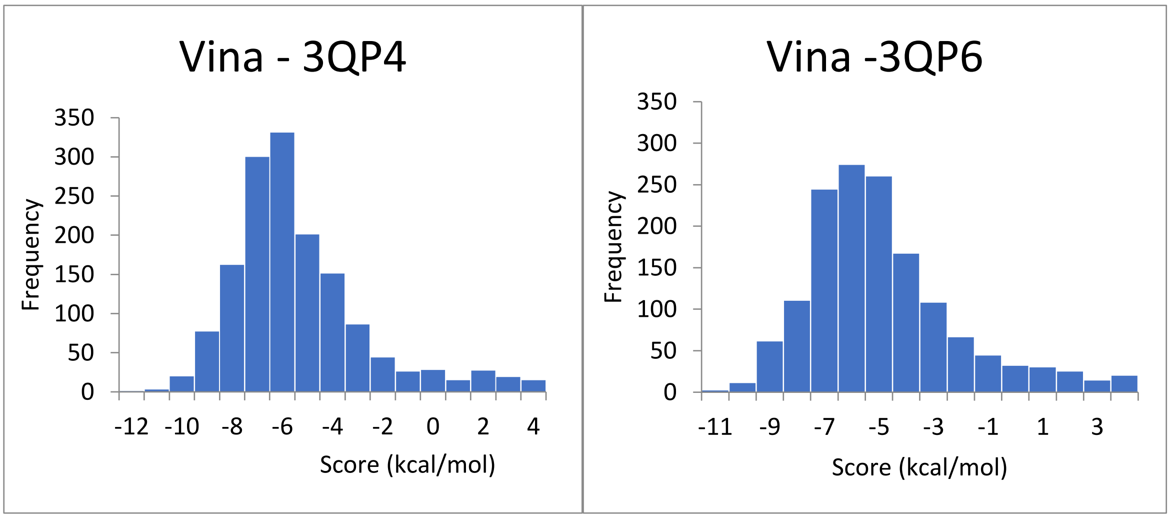


Figure 23 - Histograms for the virtual screening of ZINC/FDA Approved compounds database using AutoDock Vina for both 3QP4 and 3QP6.

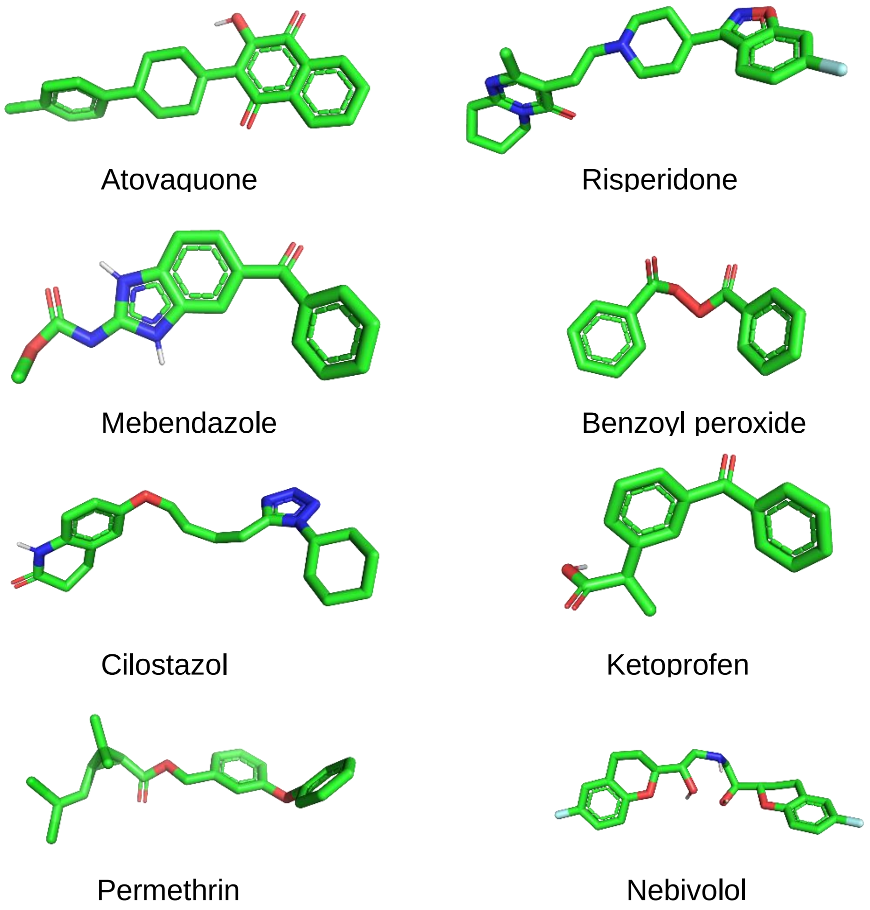
Atovaquone is the best ranked molecule for both 3QP4 and 3QP6 target. It is also important to stress that there are ten compounds present in the 25 best ranked molecules in both structures. One molecule which yielded a high score in both structures is the best ranked molecule in both virtual screenings, Atovaquone. The other nine molecules with high ranking using both structures are Risperidone, Mebendazole, Paliperidone, Permethrin, Ketoprofen, Cilostazol, Benzoyl peroxide, Iloperidone and Nebivolol.

All molecules in the top 25 yielded higher scores than those obtained during the redocking step and the crossdocking with the native ligands. A comparison between the scores of the 1st and 25th best ranked molecules for each target, and the corresponding redocking and crossdocking values is presented on table 28. This indicates that these new molecules obtained from the VS may have higher affinity to CviR than its specific ligands.

Table 28 - Comparison between the scores for the 1st and 25th best ranked molecules, obtained from ZINC/FDA database using AutoDock Vina, and the corresponding redocking and crossdocking values.

3QP4		3QP6	
Atovaquone	-12.1	Atovaquone	-11.9
Vemurafenib	-9.8	Ziprasidone	-9.6
Re-docking	-7.8	Re-docking	-7.8
Crossdocking	-8.2	Crossdocking	-8.3

Some of the highest scored molecules docked poses are shown in figure 24 and more information about these drugs is provided on table 29.



Atovaquone

Risperidone

Mebendazole

Benzoyl peroxide

Cilostazol

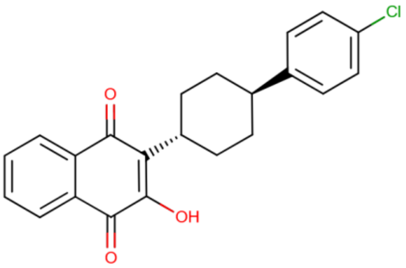
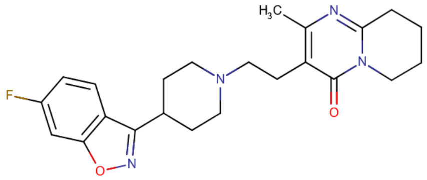
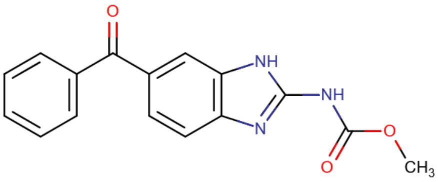
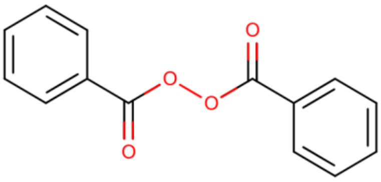
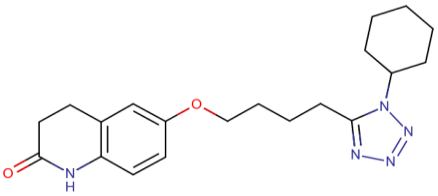
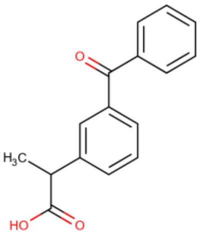
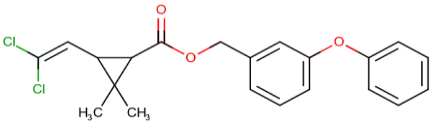
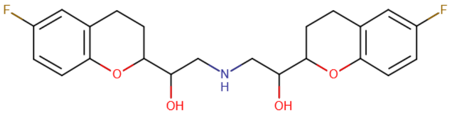
Ketoprofen

Permethrin

Nebivolol

Figure 24 - Some of the best ranked compounds, obtained from the ZINC/FDA database using AutoDock Vina, in both 3QP4 and 3QP6 targets.

Table 29 – Information on some of the best ranked compounds, obtained of from the ZINC/FDA Approved database using Autodock Vina, in both 3QP4 and 3QP6 targets.

Molecule	Description	2D Structure
Atovaquone	Atovaquone is an analog of ubiquinone, that has antimicrobial and antipneumocystis activity. It can selectively affect mitochondrial electron transport and parallel processes such as ATP and pyrimidine biosynthesis in atovaquone-responsive parasites. Its currently used in antimalarial protocols[178].	
Risperidone	Risperidone is an atypical antipsychotic with a high affinity towards 5-hydroxytryptamine (5-HT)$_{2A}$, dopamine D$_2$ and α$_1$- and α$_2$-adrenergic receptors[179].	
Mebendazole	Mebendazole is a broad spectrum anthelmintic. It binds to the colchicine-sensitive site of tubulin, inhibiting its polymerization or assembly into microtubules[180].	
Benzoyl peroxide	Benzoyl peroxide is used for topical acne therapy. This drug has broad-spectrum bactericidal activity due to its powerful oxidizing activity[181].	
Cilostazol	Cilostazol and multiple of its metabolites are cyclic AMP phosphodiesterase III inhibitors. They originate in an increase in cAMP in platelets and blood vessels, which leads to an inhibition of platelet aggregation and vasodilation.[182]	
Ketoprofen	Ketoprofen is a nonsteroidal anti-inflammatory drug with analgesic properties. These properties are due to the suppression of prostaglandin synthesis through cyclo-oxygenase inhibition[183].	
Permethrin	Permethrin is active against a severs pests including lice, fleas, and other arthropods. This drug acts on the nerve cell membrane, disrupting the sodium channel current. This causes delayed repolarization and paralysis of the pests[184].	
Nebivolol	Nebivolol is a cardioselective lipophilic beta-blocker that decreases vascular resistance, increases stroke volume and cardiac output, and does not affect left ventricular function[185].	

The results for the virtual screening of the ZINC/FDA database using GOLD/CHEMPLP scoring function are presented on table 30. Histograms for these virtual screening simulations are presented in figure 25.

Table 30 - Results of the virtual screening on the ZINC/FDA database using the GOLD/CHEMPLP procedure.

Virtual Screening Results					
3QP4			3QP6		
Rank	Ligand	Score	Rank	Ligand	Score
1	Montelukast	116.32	1	Iloprost	111.30
2	Vitamin K1	108.73	2	Montelukast	110.21
3	Glycerol phenylbutyrate	108.65	3	Glycerol phenylbutyrate	110.08
4	Bazedoxifene	107.61	4	Bazedoxifene	108.90
5	Pimozide	106.69	5	Atazanavir	108.20
6	Cobicistat	106.19	6	Travoprost	106.51
7	Imatinib	105.99	7	Salmeterol	106.14
8	Raloxifene	105.50	8	Pimozide	105.31
9	Ketoconazole	104.70	9	Iloperidone	105.20
10	Nefazodone	103.83	10	Raloxifene	104.50
11	Zafirlukast	103.23	11	Vitamin K1	103.95
12	Salmeterol	102.34	12	Nonoxynol-9	103.46
13	Silodosin	102.00	13	Nefazodone	103.27
14	Vilanterol	101.24	14	Mirabegron	102.51
15	Iloprost	100.83	15	Lapatinib	101.93
16	Itraconazole	100.28	16	Imatinib	101.78
17	Iloperidone	100.14	17	Deferoxamine	101.31
18	Dronedarone	100.02	18	Ketoconazole	101.30
19	Orlistat	98.21	19	Dabigatran etexilate	100.56
20	Fosinopril	98.10	20	Tirofiban	100.26
21	Nintedanib	97.62	21	Cetylpyridinium	100.10
22	Posaconazole	97.48	22	Orlistat	99.79
23	Permethrin	97.30	23	Cobicistat	99.69
24	Pimavanserin	97.07	24	Posaconazole	99.43
25	Thonzonium	97.02	25	Itraconazole	98.99

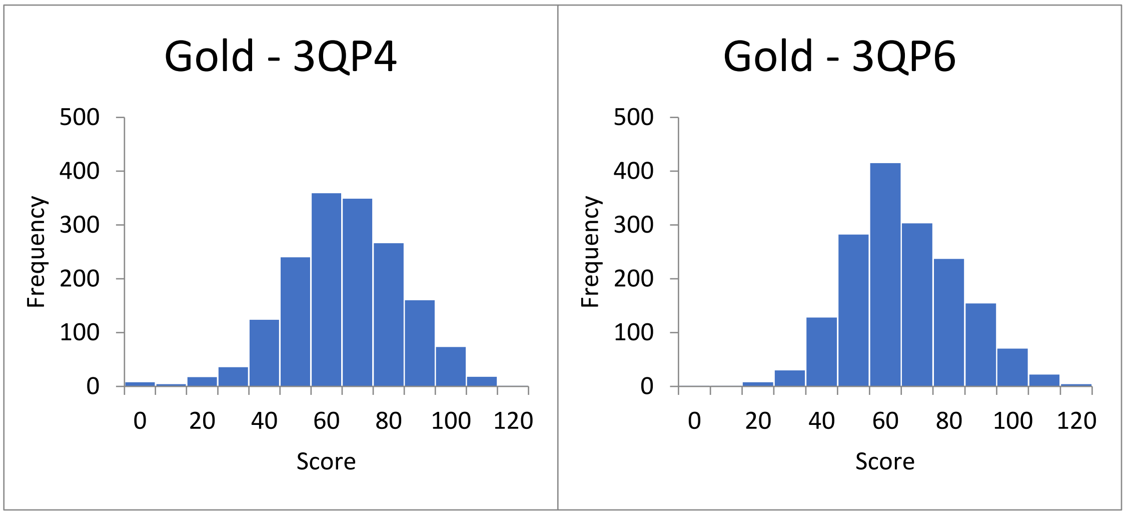


Figure 25 - Histograms for the virtual screening of ZINC/FDA database using the GOLD/CHEMPLP procedure.

The two best ranked molecules are Montelukast in 3QP4 and Iloprost in 3QP6. Both molecules gave high ranking in both structures. As before, there are other molecules that generated high scores with both structures. These molecules were Glycerol phenylbutyrate, Phylloquinone, Bazedoxifene, Raloxifene, Ketoconazole, Nefazodone and Salmeterol. All molecules in the top 25 generated higher scores than those obtained during the redocking step and the crossdocking with the native ligands. A comparison between the scores of the 1st and 25th best ranked molecules, for each target structure, and the corresponding redocking and crossdocking values is presented on table 31. This indicates that these new molecules, obtained from the VS simulations may have higher affinity to CviR than its specific ligands. Some of these molecules are presented in figure 26, and a description of their current use is provided on table 32.

Table 31 - Comparison between the scores for the 1st and 25th best ranked molecules, obtained from ZINC/FDA database using the GOLD/CHEMPLP procedure, and the corresponding redocking and crossdocking values.

3QP4		3QP6	
Montelukast	116.32	Iloprost	111.30
Thonzonium	97.02	Itraconazole	98.99
Re-docking	69.70	Re-docking	71.70
Crossdocking	87.20	Crossdocking	89.30

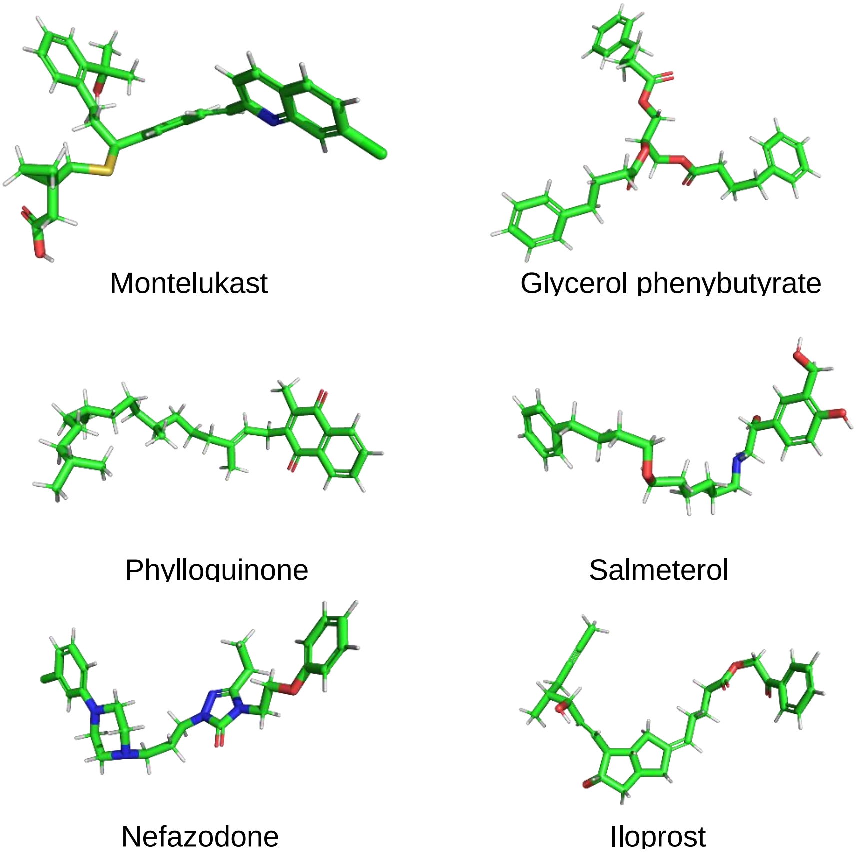
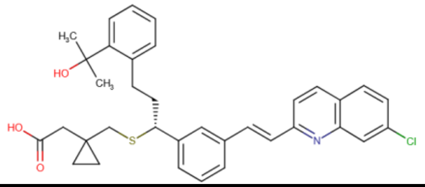
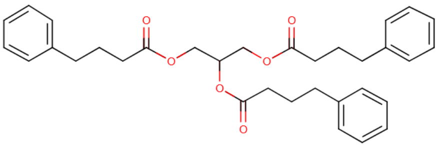
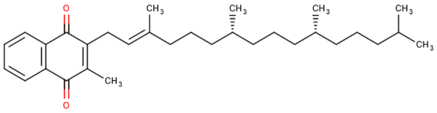
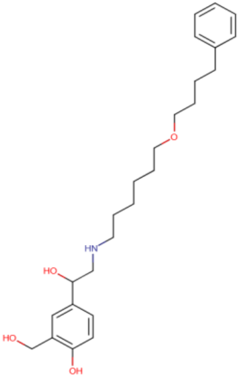
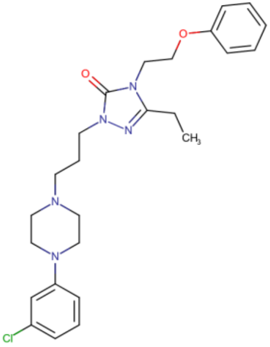
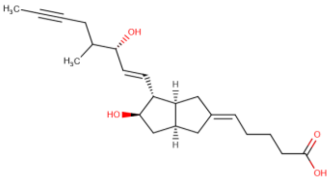


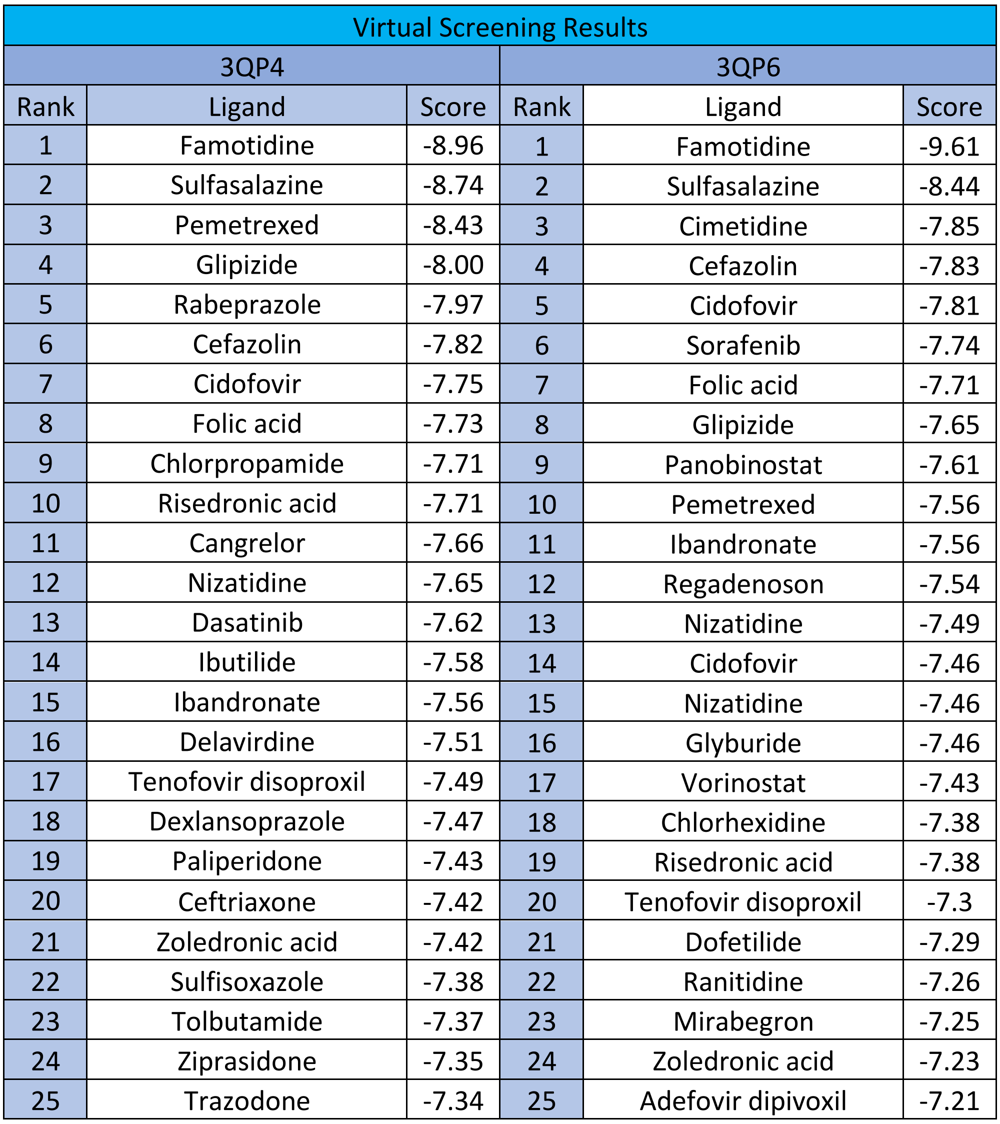
Figure 26 - Some of the best ranked docked compounds, obtained from the ZINC/FDA database using the GOLD/CHEMPLP procedure, in both 3QP4 and 3QP6 targets.

Table 32 - Information on some of the best ranked compounds, obtained from the ZINC/FDA database using the GOLD/CHEMPLP procedure, in both 3QP4 and 3QP6 targets.

Molecule	Description	2D Structure
Montelukast	Montelukast is a potent selective antagonist of the leukotriene D4 (LTD4) receptor that was developed as an oral treatment for adults and children with asthma[186].	
Glycerol phenybutyrate	Glycerol phenylbutyrate works as a prodrug. Phenylacetic acid, the major metabolite, binds to nitrogen by conjugating with glutamine through acetylation in the liver and kidneys to form phenylacetylglutamine, which is excreted by the kidneys. This provides an alternative nitrogen elimination pathway for people with Urea cycle disorders[187].	
Phylloquinone	Phylloquinone, or vitamin K1, is an essential cofactor for the gamma-carboxylase enzymes which catalyse the posttranslational gamma-carboxylation of glutamic acid residues in inactive hepatic precursors of coagulation factors II, VII, IX and X[188].	
Salmeterol	Salmeterol is an inhaled long-acting selective β_2-adrenergic receptor agonist that is currently prescribed for the treatment of asthma and chronic obstructive pulmonary disease[189].	
Nefazodone	Nefazodone is an antidepressant which acts as an antagonist for type 2 serotonin (5-HT$_2$) post-synaptic receptors and moderately inhibits serotonin and noradrenaline reuptake[190].	
Iloprost	Iloprost is a synthetic analogue of prostacyclin PGI$_2$ which dilates systemic and pulmonary arterial vascular beds. It is used for the treatment of pulmonary arterial hypertension[191].	

The results of the virtual screening of the ZINC/FDA database using LeDock are displayed in on table 33. Histograms for these virtual screening simulations are presented in figure 27.

Table 33 - Results of the virtual screening of the ZINC/FDA database using LeDock Scores are expressed in kcal/mol.



Virtual Screening Results					
3QP4			3QP6		
Rank	Ligand	Score	Rank	Ligand	Score
1	Famotidine	-8.96	1	Famotidine	-9.61
2	Sulfasalazine	-8.74	2	Sulfasalazine	-8.44
3	Pemetrexed	-8.43	3	Cimetidine	-7.85
4	Glipizide	-8.00	4	Cefazolin	-7.83
5	Rabeprazole	-7.97	5	Cidofovir	-7.81
6	Cefazolin	-7.82	6	Sorafenib	-7.74
7	Cidofovir	-7.75	7	Folic acid	-7.71
8	Folic acid	-7.73	8	Glipizide	-7.65
9	Chlorpropamide	-7.71	9	Panobinostat	-7.61
10	Risedronic acid	-7.71	10	Pemetrexed	-7.56
11	Cangrelor	-7.66	11	Ibandronate	-7.56
12	Nizatidine	-7.65	12	Regadenoson	-7.54
13	Dasatinib	-7.62	13	Nizatidine	-7.49
14	Ibutilide	-7.58	14	Cidofovir	-7.46
15	Ibandronate	-7.56	15	Nizatidine	-7.46
16	Delavirdine	-7.51	16	Glyburide	-7.46
17	Tenofovir disoproxil	-7.49	17	Vorinostat	-7.43
18	Dexlansoprazole	-7.47	18	Chlorhexidine	-7.38
19	Paliperidone	-7.43	19	Risedronic acid	-7.38
20	Ceftriaxone	-7.42	20	Tenofovir disoproxil	-7.3
21	Zoledronic acid	-7.42	21	Dofetilide	-7.29
22	Sulfisoxazole	-7.38	22	Ranitidine	-7.26
23	Tolbutamide	-7.37	23	Mirabegron	-7.25
24	Ziprasidone	-7.35	24	Zoledronic acid	-7.23
25	Trazodone	-7.34	25	Adefovir dipivoxil	-7.21

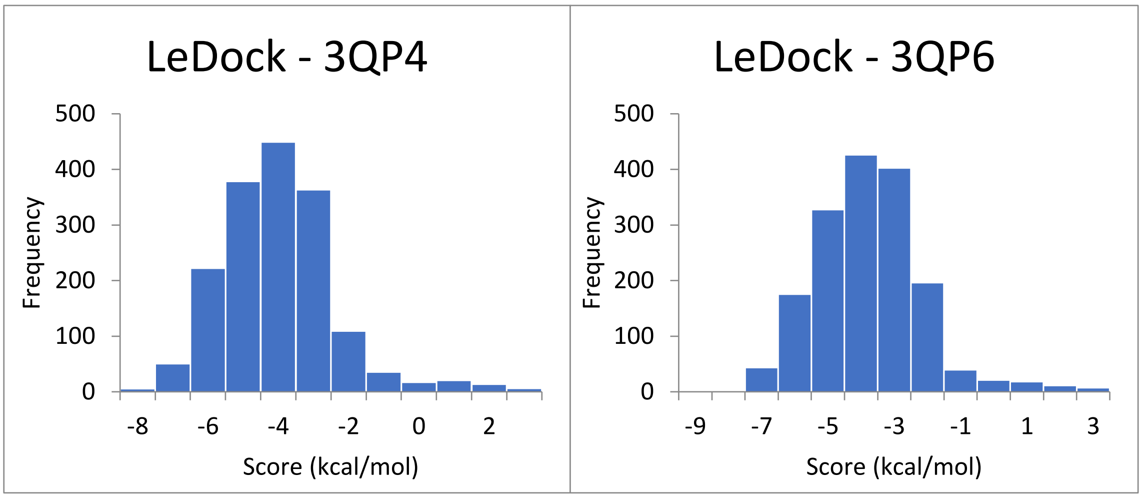


Figure 27 – Histograms for the virtual screening of ZINC/FDA database using LeDock

In both structures the best scored molecule is Famotidine, followed by Sulfasalazine. There are nine others, which generated high scores in both structures. These compounds are Pemetrexed, Glipizide, Cefazolin, Cidofovir, Folic acid, Risedronic acid, Nizatidine, Tenofovir disoproxil and Zoledronic acid. As on the virtual screenings using the previous two molecular docking procedures, the top 25 ranked molecules yielded higher scores than those generated on the redocking and crossdocking experiments (see table 34). Some of these molecules are shown on figure 28 and more information about them is provided on table 35.

Table 34 - Comparison between the scores for the 1st and 25th best ranked molecules, obtained from ZINC/FDA database using LeDock.

3QP4		3QP6	
Famotidine	-8.96	Famotidine	-9.61
Trazodone	-7.34	Adefovir dipivoxil	-7.21
Re-docking	-6.04	Re-docking	-5.87
Crossdocking	-6.92	Crossdocking	-7.00

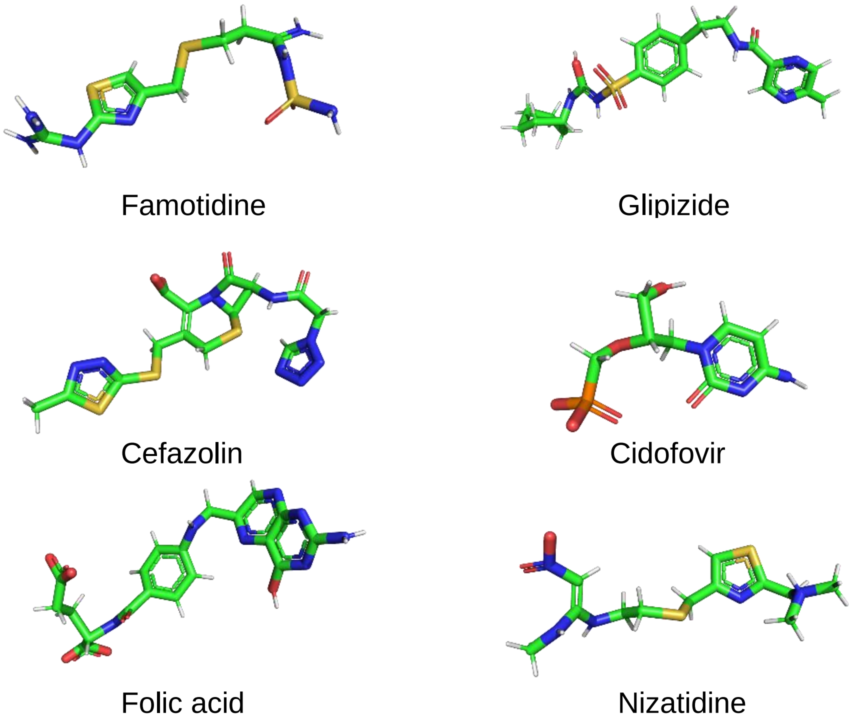
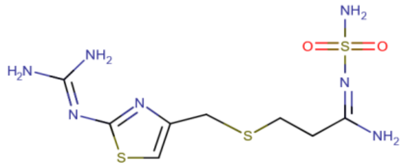
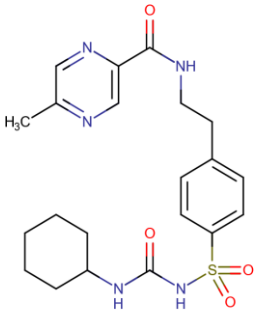
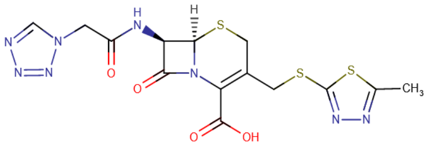
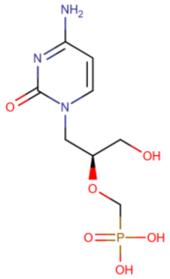
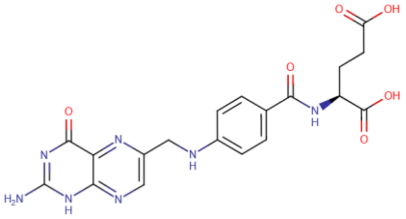
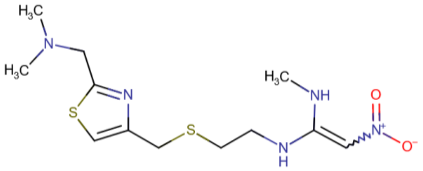


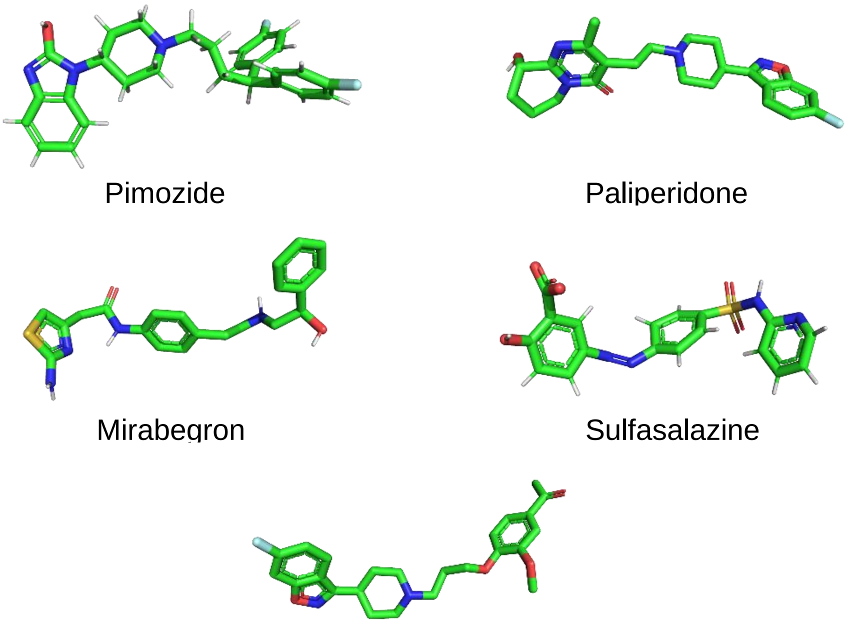
Figure 28- Some of the best ranked docked compounds, obtained from the ZINC/ FDA database obtained using LeDock, in both 3QP4 and 3QP6 targets.

Table 35 - Information on some of the best ranked compounds, obtained from the ZINC/FDA database using LeDock, in both 3QP4 and 3QP6 targets.

Molecule	Description	2D Structure
Famotidine	Famotidine is a competitive histamine H_2 receptor antagonist. It is used for gastrointestinal conditions related to acid secretion, such as gastric ulcers[192].	
Glipizide	Glipizide is part of the sulfonylurea drug class. This drug acts by stimulating the pancreatic β-cells to secrete insulin. This is done by binding to receptors that block the potassium ATP-dependent channels. This depolarises the cell leading to insulin exocytosis[193].	
Cefazolin	Cefazolin is a cephalosporin analog with broad-spectrum antibiotic action. Its effect is achieved by binding to specific penicillin-binding proteins inside the bacterial cell wall, inhibiting the synthesis of the cell wall[194].	
Cidofovir	Cidofovir is an antiviral drug used for the treatment of cytomegalovirus retinitis in patients diagnosed with AIDS. It acts by inhibiting the viral DNA polymerase[195].	
Folic acid	Folic acid is used to treat tetrahydrofolic acid and vitamin B12 deficiencies. It is converted to tetrahydrofolic acid by dihydrofolate reductase[122].	
Nizatidine	Nizatidine is a histamine H_2 receptor antagonist used for the treatment of duodenal ulcers[196].	

There are a reduced number of molecules that yielded high scores using more than one molecular docking. Two molecules (Pimozide and Iloperidone) presented high scores using both AutoDock Vina and GOLD/CHEMPLP procedures. Three other molecules (Paliperidone, Mirabegron and Iloperidone) had high scores using AutoDock

Vina and LeDock. These molecules, presented in figure 29, are more likely to have high affinity towards CviR. Further information on these compounds is presented in table 36.



Pimozide

Paliperidone

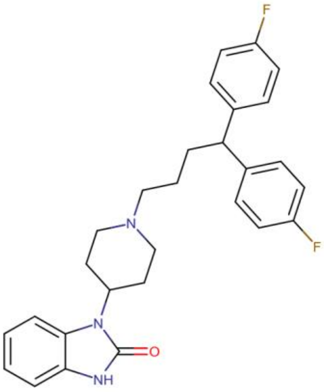
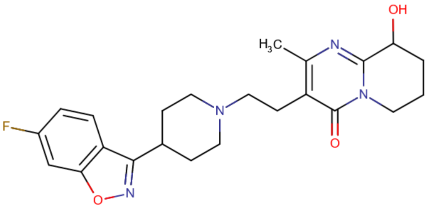
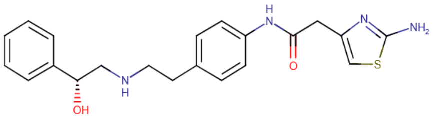
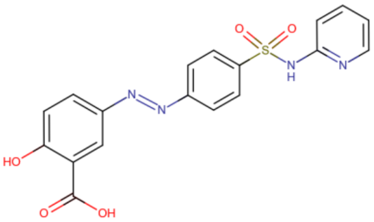
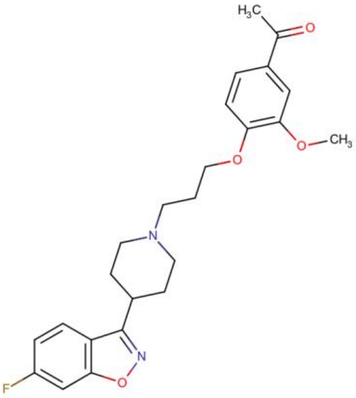
Mirabegron

Sulfasalazine

Iloperidone

Figure 29 – Molecules that generated high scores in multiple molecular docking procedures.

Table 36 - Further information on the compounds that generated high scores in multiple molecular docking procedures.

Molecule	Description	2D Structure
Pimozide	Pimozide is used as an antipsychotic agent and for the suppression of vocal and motor tics in patients with Tourette syndrome. Although its exact mechanism of action is unknown, it is thought that it inhibits the dopamine D_2 receptor[197].	
Paliperidone	Palperidone is a metabolite of risperidone, now also used as an antipsychotic. The mechanism of action is unknown but it is likely to act via a similar pathway to risperidone[198].	
Mirabegron	Mirabegron is selective β_3-Adrenoceptor Agonist. It is used for the treatment of symptoms of overactive bladder[199].	
Sulfasalazine	Sulfasalazine is an anti-inflammatory drug used for the treatment of ulcerative colitis and rheumatoid arthritis. Its activity is believed to be due to its metabolites 5-aminosalicylic acid and sulfapyridine[200].	
Iloperidone	Iloperidone is an antipsychotic for the treatment of schizophrenia symptoms. It shows high affinity and maximal receptor occupancy for dopamine D2 receptors in the caudate nucleus and putamen of the brains of schizophrenic patients[201].	

4.4.2.2 Mu.Ta.Lig Virtual Chemotheca

The results of the virtual screening of the Chemotecha database, obtained using Autodock Vina, are presented on table 37 and figure 30.

Table 37 - The 25 best ranked molecules of obtained by virtual screening on the Chemotheca database using Autodock Vina. Scores are expressed in kcal/mol.

Virtual Screening Results					
3QP4			3QP6		
Rank	Ligand	Score	Rank	Ligand	Score
1	CMLDID18049	-12.1	1	CMLDID40723	-12.1
2	CMLDID40723	-12.0	2	CMLDID2574	-12.0
3	CMLDID28191	-11.9	3	CMLDID18049	-11.8
4	CMLDID42504	-11.9	4	CMLDID29755	-11.8
5	CMLDID21539	-11.8	5	CMLDID59940	-11.8
6	CMLDID2574	-11.8	6	CMLDID60625	-11.8
7	CMLDID44495	-11.8	7	CMLDID12797	-11.7
8	CMLDID49025	-11.8	8	CMLDID53665	-11.7
9	CMLDID56771	-11.8	9	CMLDID9445	-11.7
10	CMLDID18431	-11.7	10	CMLDID10171	-11.6
11	CMLDID20345	-11.7	11	CMLDID15258	-11.6
12	CMLDID22811	-11.7	12	CMLDID28191	-11.6
13	CMLDID35301	-11.7	13	CMLDID42504	-11.6
14	CMLDID39041	-11.7	14	CMLDID43475	-11.6
15	CMLDID53028	-11.7	15	CMLDID16034	-11.5
16	CMLDID61796	-11.7	16	CMLDID17663	-11.5
17	CMLDID13405	-11.6	17	CMLDID19001	-11.5
18	CMLDID22732	-11.6	18	CMLDID22279	-11.5
19	CMLDID32434	-11.6	19	CMLDID22810	-11.5
20	CMLDID33654	-11.6	20	CMLDID34358	-11.5
21	CMLDID48212	-11.6	21	CMLDID18431	-11.4
22	CMLDID50158	-11.6	22	CMLDID23260	-11.4
23	CMLDID610	-11.6	23	CMLDID24236	-11.4
24	CMLDID12797	-11.5	24	CMLDID36243	-11.4
25	CMLDID1363	-11.5	25	CMLDID49245	-11.4

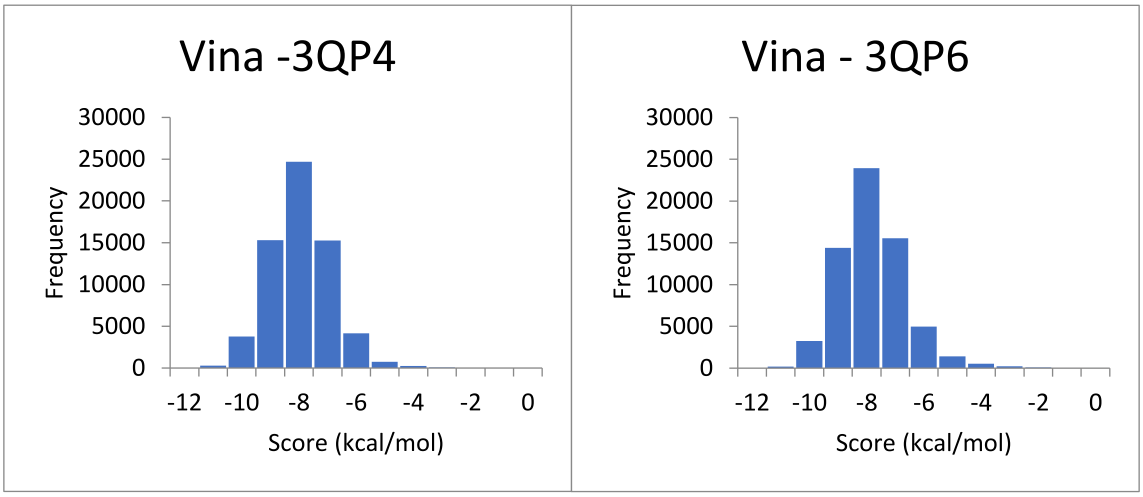


Figure 30 - Histograms for the virtual screening of the Chemotheca database using Autodock Vina

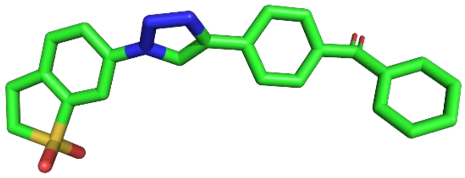
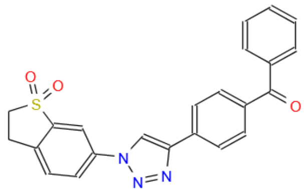
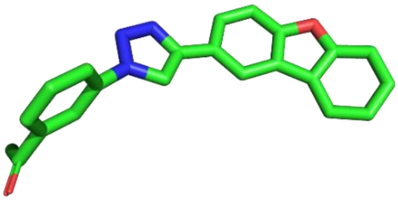
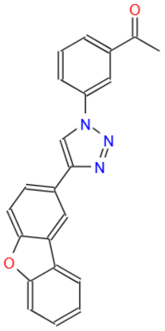
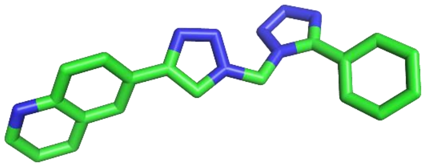
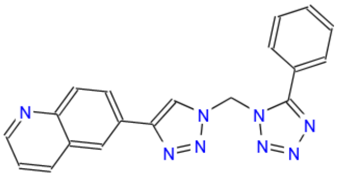
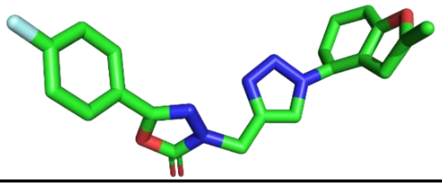
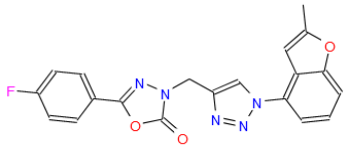
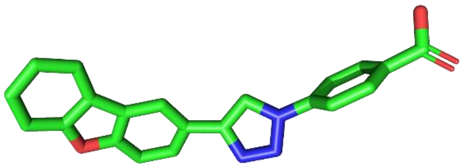
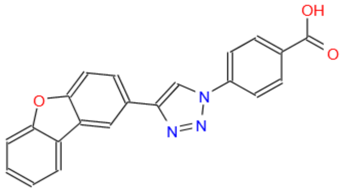
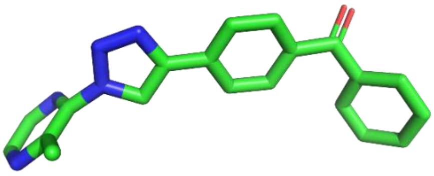
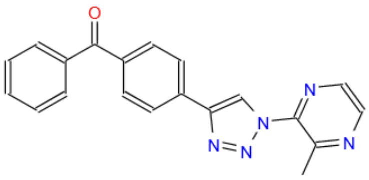
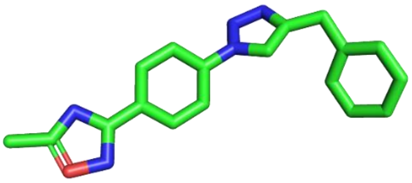
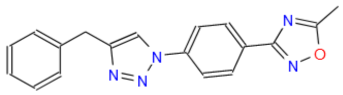
All the 25 best ranked molecules generated much higher scores than those obtained on the redocking and crossdocking studies (see table 38).

Table 38 - Comparison between the scores for the 1st and 25th best ranked molecules, obtained from Chemotheca database using AutoDock Vina, and the corresponding redocking and crossdocking values.

3QP4		3QP6	
CMLDID18049	-12.10	CMLDID40723	-12.10
CMLDID1363	-11.50	CMLDID49245	-11.40
Re-docking	-7.80	Re-docking	-7.80
Crossdocking	-8.20	Crossdocking	-8.30

There were seven molecules that ranked in the top 25 in both structures. These molecules are presented on table 39.

Table 39 - Best ranked compounds, obtained from the Chemotheca database using AutoDock Vina, in both 3QP4 and 3QP6 targets.

Molecule	Docked pose	2D Structure
CMLDID18049		
CMLDID40723		
CMLDID28191		
CMLDID42504		
CMLDID2574		
CMLDID18431		
CMLDID12797		

The results for the virtual screening of the Chemotheca database using GOLD/CHEMPLP are presented on table 40 and figure 31.

Table 40 - The 25 best ranked molecules obtained by virtual screening on the Chemotheca database using the GOLD/CHEMPLP procedure.

Virtual Screening Results					
3QP4			3QP6		
Rank	Ligand	Score	Rank	Ligand	Score
1	CMLDID5450	116.23	1	CMLDID5450	124.35
2	CMLDID11709	115.42	2	CMLDID17434	123.05
3	CMLDID23812	115.22	3	CMLDID29373	121.98
4	CMLDID17434	115.22	4	CMLDID23842	121.32
5	CMLDID6261	114.34	5	CMLDID50715	118.84
6	CMLDID18556	114.25	6	CMLDID28747	118.66
7	CMLDID64050	113.19	7	CMLDID38590	118.60
8	CMLDID38590	112.36	8	CMLDID23812	118.25
9	CMLDID27098	112.28	9	CMLDID44663	116.66
10	CMLDID57782	112.24	10	CMLDID61216	116.18
11	CMLDID23842	112.13	11	CMLDID23215	116.10
12	CMLDID62223	112.01	12	CMLDID11154	115.62
13	CMLDID25998	111.94	13	CMLDID14675	115.50
14	CMLDID20688	111.92	14	CMLDID29578	115.39
15	CMLDID33851	111.88	15	CMLDID16134	115.13
16	CMLDID11293	111.86	16	CMLDID44012	114.60
17	CMLDID31711	111.78	17	CMLDID20688	114.57
18	CMLDID54588	111.77	18	CMLDID3837	114.44
19	CMLDID19295	111.72	19	CMLDID54632	114.22
20	CMLDID3426	111.58	20	CMLDID20544	114.11
21	CMLDID49778	111.58	21	CMLDID17058	114.04
22	CMLDID12819	111.51	22	CMLDID18631	113.94
23	CMLDID46450	111.47	23	CMLDID13117	113.84
24	CMLDID6049	111.43	24	CMLDID58653	113.63
25	CMLDID44834	111.40	25	CMLDID31711	113.61

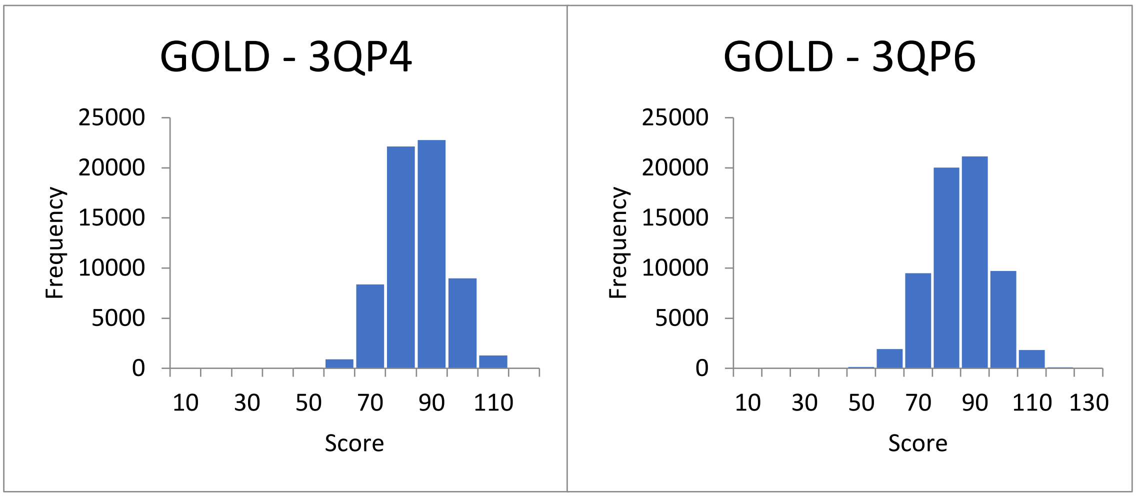


Figure 31 - Histograms for the virtual screening of Chemotheca database using the GOLD/CHEMPLP procedure

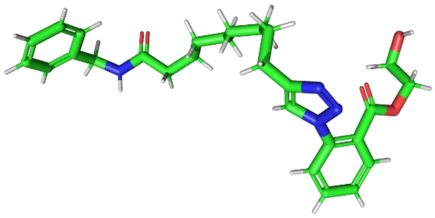
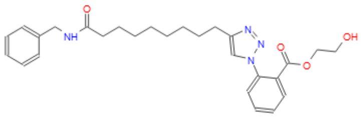
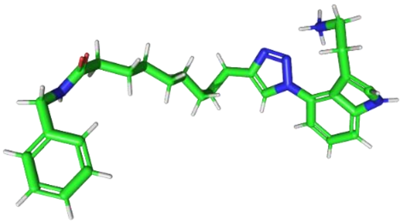
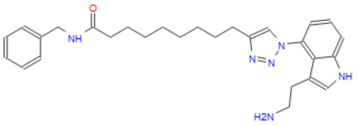
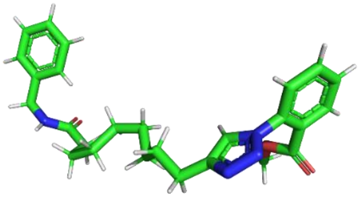
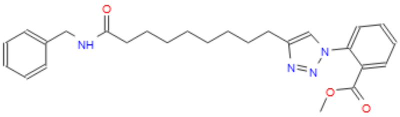
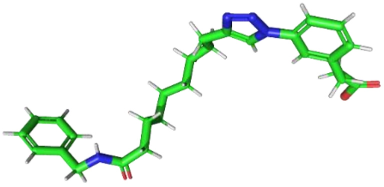
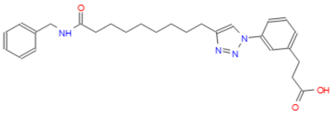
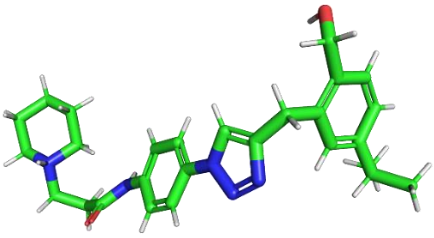
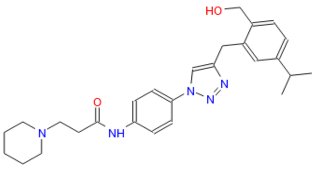
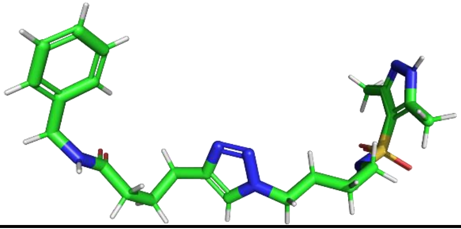
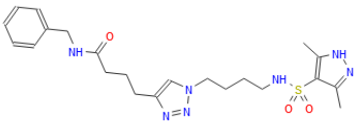
As before, the 25 best ranked molecules generated higher scores than those obtained during the redocking and crossdocking steps of this work (see table 41).

Table 41 - Comparison between the scores for the 1st and 25th best ranked molecules, obtained from the Chemtheca database using the GOLD/CHEMPLP procedure, and the corresponding redocking and crossdocking values.

3QP4		3QP6	
CMLDID5450	116.23	CMLDID5450	124.35
CMLDID44834	111.40	CMLDID31711	113.61
Re-docking	69.70	Re-docking	71.70
Crossdocking	87.20	Crossdocking	89.30

There were 6 molecules which placed in the top 25 in both virtual screenings using the GOLD/CHEMPLP procedure. Their 2D structures and docked poses are provided in the table 42.

Table 42 - Best ranked compounds, obtained from the Chemotheca database using the GOLD/CHEMPLP procedure, in both 3QP4 and 3QP6 targets.

Molecule	Docked pose	2D Structure
CMLDID5450		
CMLDID17434		
CMLDID23812		
CMLDID38590		
CMLDID20688		
CMLDID31711		

The results for the virtual screening on the Chemotheca database using LeDock are presented in table 43 and figure 32.

Table 43 - The 25 best ranked molecules obtained by virtual screening on the Mu.Ta.Lig database using LeDock. Scores are expressed in kcal/mol.

Virtual Screening Results					
3QP4			3QP6		
Rank	Ligand	Score	Rank	Ligand	Score
1	CMLDID34044	-9.87	1	CMLDID35542	-10.00
2	CMLDID37602	-9.86	2	CMLDID60399	-9.93
3	CMLDID53319	-9.82	3	CMLDID56610	-9.81
4	CMLDID22194	-9.79	4	CMLDID59293	-9.77
5	CMLDID11222	-9.76	5	CMLDID34911	-9.76
6	CMLDID54235	-9.74	6	CMLDID50121	-9.71
7	CMLDID50121	-9.73	7	CMLDID53952	-9.71
8	CMLDID36452	-9.65	8	CMLDID39280	-9.70
9	CMLDID50368	-9.65	9	CMLDID29586	-9.62
10	CMLDID53952	-9.63	10	CMLDID46590	-9.58
11	CMLDID20643	-9.61	11	CMLDID34296	-9.55
12	CMLDID28193	-9.61	12	CMLDID57884	-9.53
13	CMLDID35542	-9.61	13	CMLDID63369	-9.50
14	CMLDID42086	-9.60	14	CMLDID20822	-9.49
15	CMLDID43726	-9.60	15	CMLDID30007	-9.47
16	CMLDID60399	-9.59	16	CMLDID34451	-9.47
17	CMLDID55168	-9.56	17	CMLDID22894	-9.45
18	CMLDID65218	-9.56	18	CMLDID50451	-9.45
19	CMLDID60211	-9.55	19	CMLDID59073	-9.44
20	CMLDID62031	-9.55	20	CMLDID16747	-9.40
21	CMLDID15435	-9.54	21	CMLDID17650	-9.39
22	CMLDID16776	-9.53	22	CMLDID46997	-9.39
23	CMLDID10746	-9.50	23	CMLDID56704	-9.38
24	CMLDID44007	-9.50	24	CMLDID22361	-9.37
25	CMLDID5331	-9.50	25	CMLDID4018	-9.37

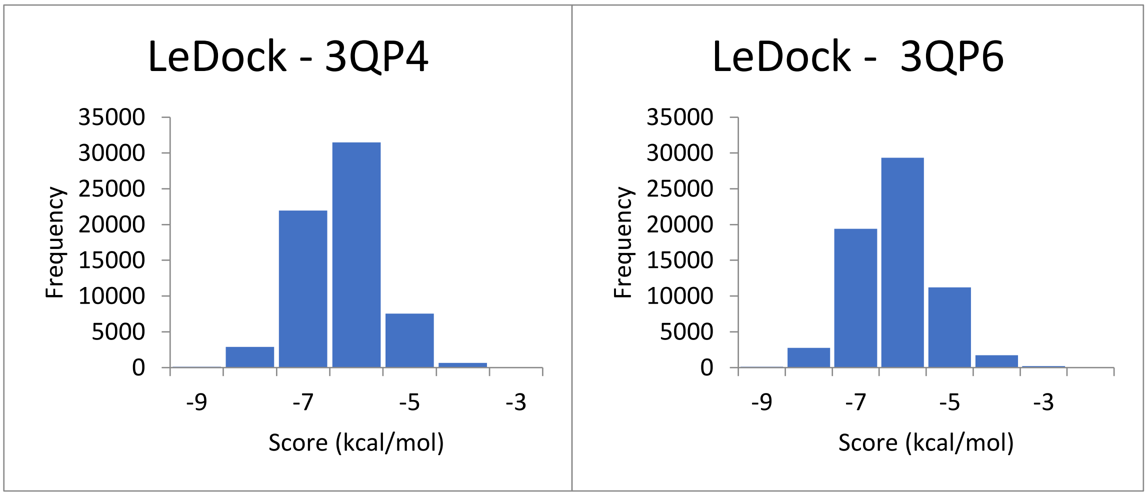


Figure 32 - Histograms for the virtual screening of Chemotheca database using LeDock

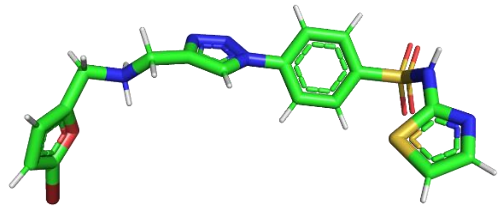
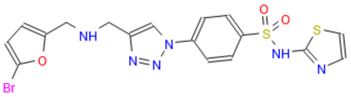
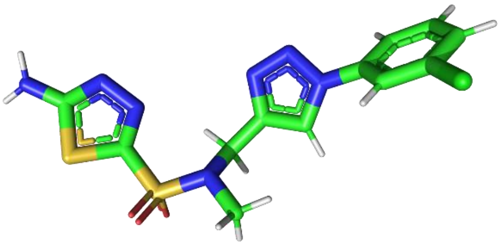
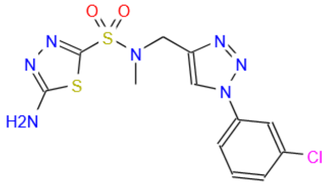
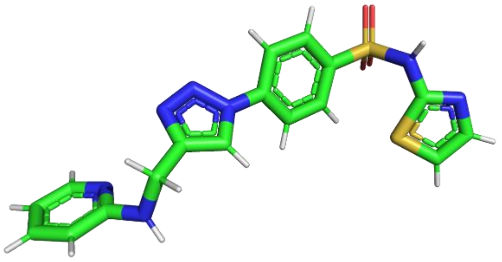
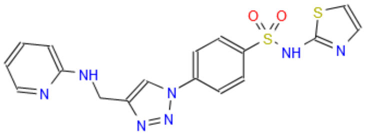
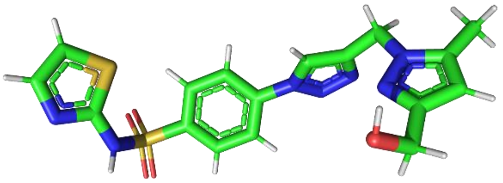
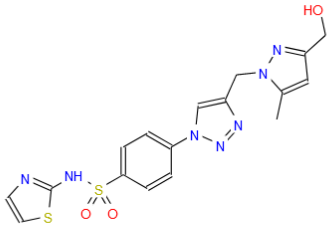
Keeping with the previous results, the best ranked molecules present much higher scores than what was generated with redocking and crossdocking.

Table 44 - Comparison between the scores for the 1st and 25th best ranked molecules, obtained from the Chemotheca database using LeDock, and the corresponding redocking and crossdocking values.

3QP4		3QP6	
CMLDID34044	-9.87	CMLDID35542	-10.00
CMLDID5331	-9.50	CMLDID4018	-9.37
Re-docking	-6.04	Re-docking	-5.87
Crossdocking	-6.92	Crossdocking	-7.00

After analysing the final results, it can be seen that there were 4 molecules which ranked in the best 25 scores on both the virtual screening using 3QP4 and 3QP6. Their 2D structures and docked poses of these molecules are available on table 45.

Table 45 - Best ranked compounds, obtained from the Chemotheca database using LeDock, in both 3QP4 and 3QP6 targets.

Molecule	Docked pose	2D Structure
CMLDID50121		
CMLDID53952		
CMLDID35542		
CMLDID60399		

4.4.3 Overall

Figures 33 and 3 summarize all virtual screening simulations performed with 3QP4 and 3QP6 respectively, comparing them with the scores of known active molecules.

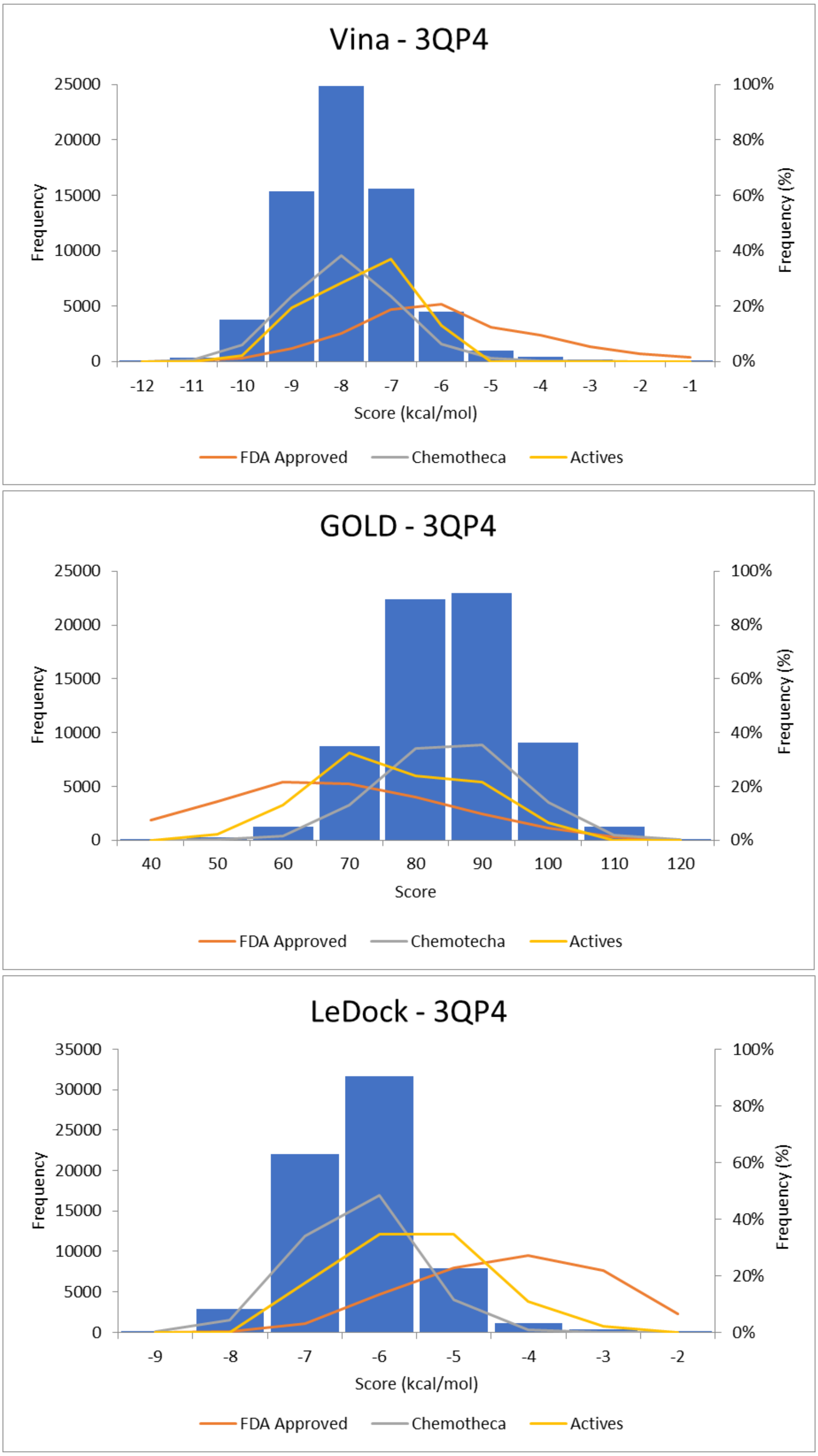
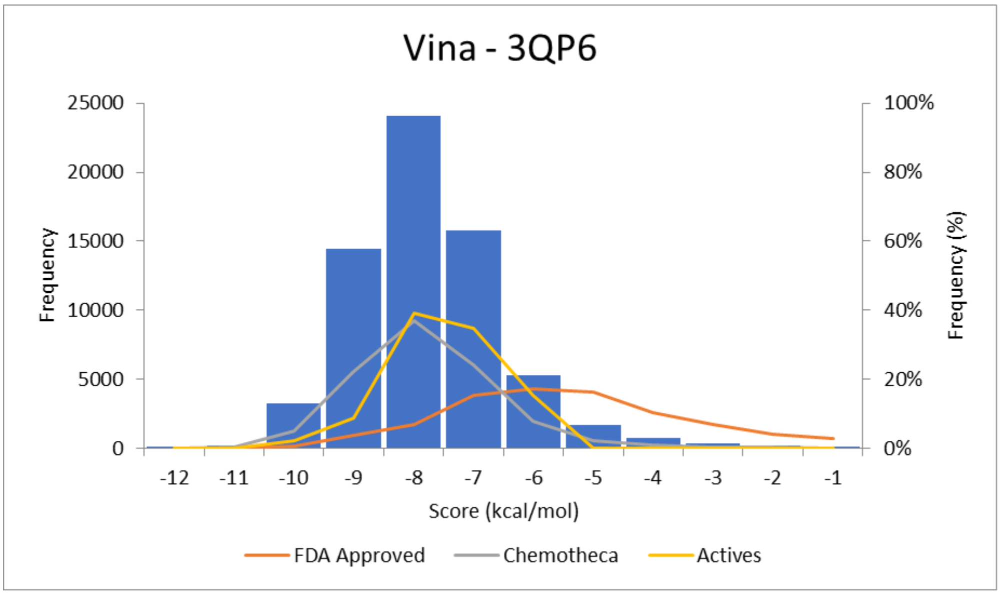
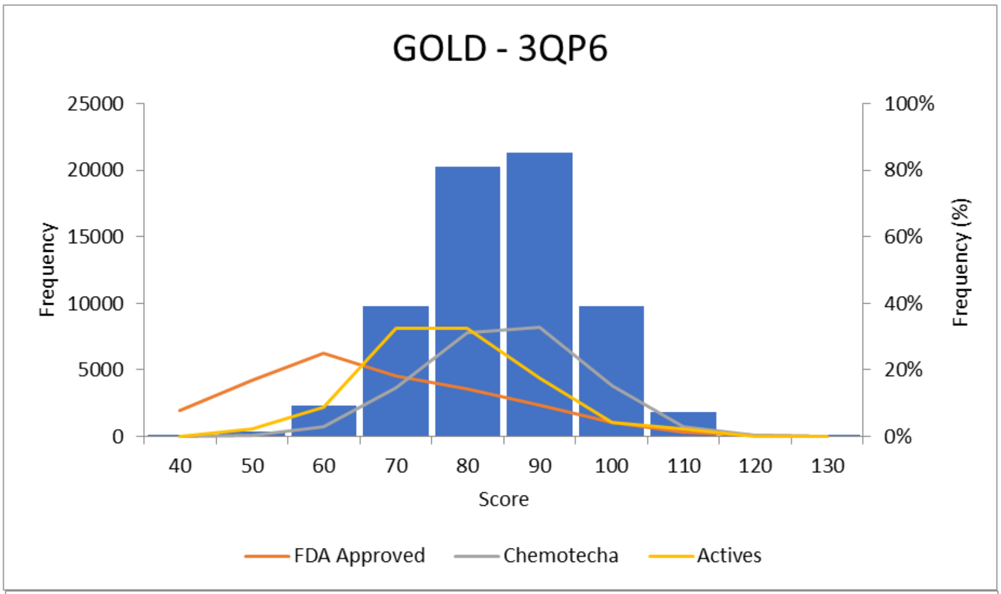


Figure 33 - Histograms all screened molecule on 3QP4 using each molecular docking program. The lines represent the distribution (in percentage) of each database and the known active molecules.





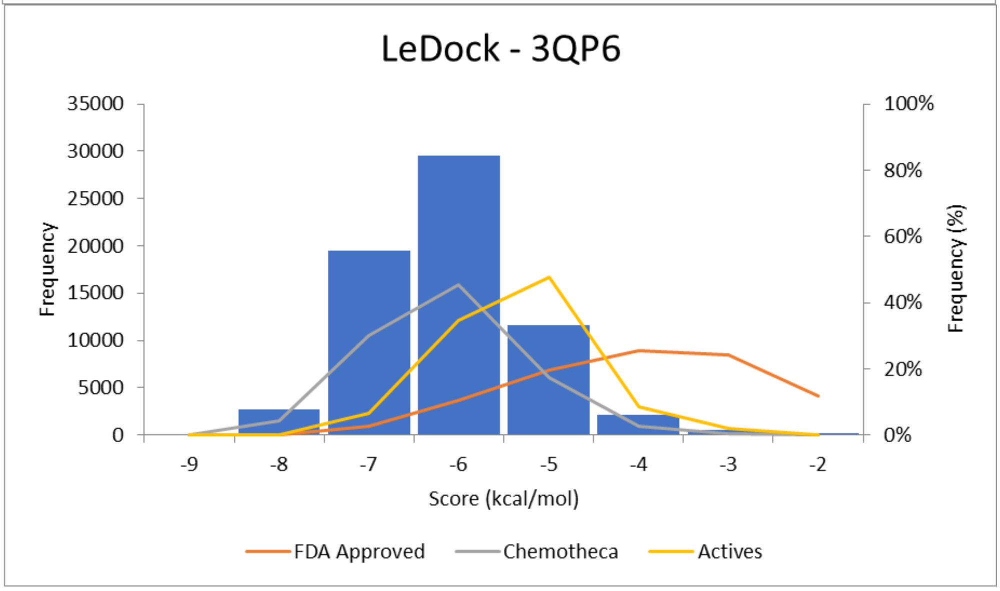


Figure 34 - Histograms all screened molecule on 3QP6 using each molecular docking program. The lines represent the distribution (in percentage) of each database and the known active molecules.

These results show that Chemotheca was the database which yielded the highest percentage of high scoring compounds. Even though the ZINC/FDA virtual screening did not generate scores as high as Chemotheca, it still resulted in multiple molecules with higher scores than the best ranked known active molecules. Using Autodock Vina, the Chemotheca virtual screenings generated over 2000 molecules with a higher score then the best scored known active, while the ZINC/FDA database generated over 10 molecules of this type. With GOLD, Chemotheca possessed over 1000 molecules with a score higher than the best scored active and the FDA Approved database virtual screening resulted in over 20 molecules. Finally, using LeDock, Chemotheca yielded over 10000 molecules with higher scores than the best scored known active and there were over 20 ZINC/FDA compounds with similar performances. Considering the total number of molecules of each database (1657 for ZINC/FDA and 64804 for Chemotheca) this difference on the number of molecules with the highest scores was expected. With this in mind, both databases had a positive performance, generating several promising molecules for a screening in a laboratory or for further computational testing.

4.4.4 Conclusions

The virtual screening experiments performed on 3QP4 and 3QP6, using two different databases and three different molecular docking procedures resulted in several promising molecules.

In order to provide a better understanding on their inhibitory potential to CviR and their capacity for hindering the quorum sensing process, the best performing compounds from each database will be further studied using more rigorous methods. Ten molecules from each database were selected for these studies. The molecules chosen from the ZINC/FDA database were Atovaquone, Famotidine, Iloprost, Mebendazole, Mirabegron, Montelukast, Paliperidone, Glycerol Phenybutyrate and Sulfasalazine. The molecules chosen from Chemotheca were CMLDID2574, CMLDID5450, CMLDID17434, CMLDID18049, CMLDID23812, CMLDID35542, CMLDID38590, CMLDID40723, CMLDID50121 and CMLDID60399. These were the top ranked molecules in virtual screening simulations using both structures of CviR or the high-placed ones by more than one molecular docking procedure.

4.5 Molecular dynamics simulations and MM/PB(GB)SA

4.5.1 Methods

4.5.1.1 Molecular dynamics simulations

The molecular dynamics simulations were performed using the Amber18 software. The selected ligands were minimized using the HF/6-31G* optimization in Gaussian[202] and the force field parameters were assigned using antechamber and LEaP programs, with RESP HF/G-31G(d) charges. The protein was described through the amber14sb force filed. The complex was embedded into a box of TIP3P water molecules, whose edges are placed at least 12 Å away from each atom of the complex. Periodic boundaries were applied, and the long-range electrostatic interactions were calculated using the particle mesh Ewald summation method. The cut-off value for the short-range electrostatic and Lennard–Jones interactions was set at 10.0 Å. All bonds involving hydrogen atoms were constrained using the SHAKE algorithm enabling the application of a 2 fs time step.

All ligand-CviR solvated complexes went through 4 minimization steps, with a maximum of 2500 cycles each. After 1250 cycles the minimization method was switched from steepest descent to conjugate gradient. In the first minimization all molecules except water molecules were restrained. In the second minimization, only hydrogen atoms were not restrained. During the third minimization only the protein backbone was restrained. Finally, for the last minimization step, there were no restraints.

Following the minimization phase, all solvated complexes were heated from 0 to 310.15 K over 50 ps. They were further equilibrated at 310.15 K during 50 ps, to stabilize the density.

Finally, the production phase was run for a total of 100 ns in an NPT ensemble at a pressure of 1 bar and a temperature of 310.15 K.

The analysis of the trajectories were carried out using the cpptraj tool[203] and its visual analysis were done using VMD.

4.5.1.2 MM/PB(GB)SA

The MM/PB(GB)SA calculations were performed using the MM/PBSA.py script available in AMBER[153]. The calculations considered the last 40 ns of the MD simulation of every complex, using an interval of 100 ps. This means that the program will use every 10th frame of simulation. In the MM/PBSA calculations, the following constants were used: ionic strength of 0.100 mol dm^{-3}, external dielectric constant of 80.0 and internal dielectric constant of 4. In the MM/GBSA calculations a salt concentration of 0.100 mol dm^{-3} was used.

The Free energy decomposition option was used in order to obtain information about the local interactions of the complex. Using per-residue decomposition, the contribution of each residue to the total free energy was estimated.

4.5.2 Results

4.5.2.1 ZINC/FDA

As stated above, the 10 molecules chosen from this database were Atovaquone, Famotidine, Iloprost, Mebendazole, Mirabegron, Montelukast, Paliperidone, Pimozide, Glycerol Phenybutyrate and Sulfasalazine. For each molecule, in complex with 3QP6, 100 ns of MD simulation was performed in order to evaluate the structural stability of the protein-ligand complex and to carry out the MM/PB(GB)SA studies. To assess the structural stability of the complex, RMSD calculations were performed for the C_α atoms of each complex and the results can be seen on table 46 and figure 35 and for the ligands, which can be seen on table 47 and figure 36. For reference, a MD simulation of C10-HSL in complex with the protein was also performed. Since this ligand was the agonist which generated the highest scores on the cross-docking studies with 3QP6, it was selected for this study. While the native ligand of the strain 12472 of *C. violaceum* is 3-hydroxy-C10-HSL, it also responds to C10-HSL[67].

Table 46 - Average RMSD values (Å) of the last 40ns of the simulation for the CviR-ligand complexes for the selected molecules from the ZINC/FDA database.

Ligand	Average RMSD	Standard Deviation
C10HSL	1.55	0.12
Atovaquone	1.78	0.19
Famotidine	1.68	0.14
Iloprost	1.72	0.08
Mebendazole	1.64	0.14
Mirabegron	1.35	0.09
Montelukast	1.50	0.09
Paliperidone	1.26	0.13
Pimozide	1.78	0.15
Glycerol Phenylbutyrate	1.29	0.10
Sulfasalazine	1.27	0.09

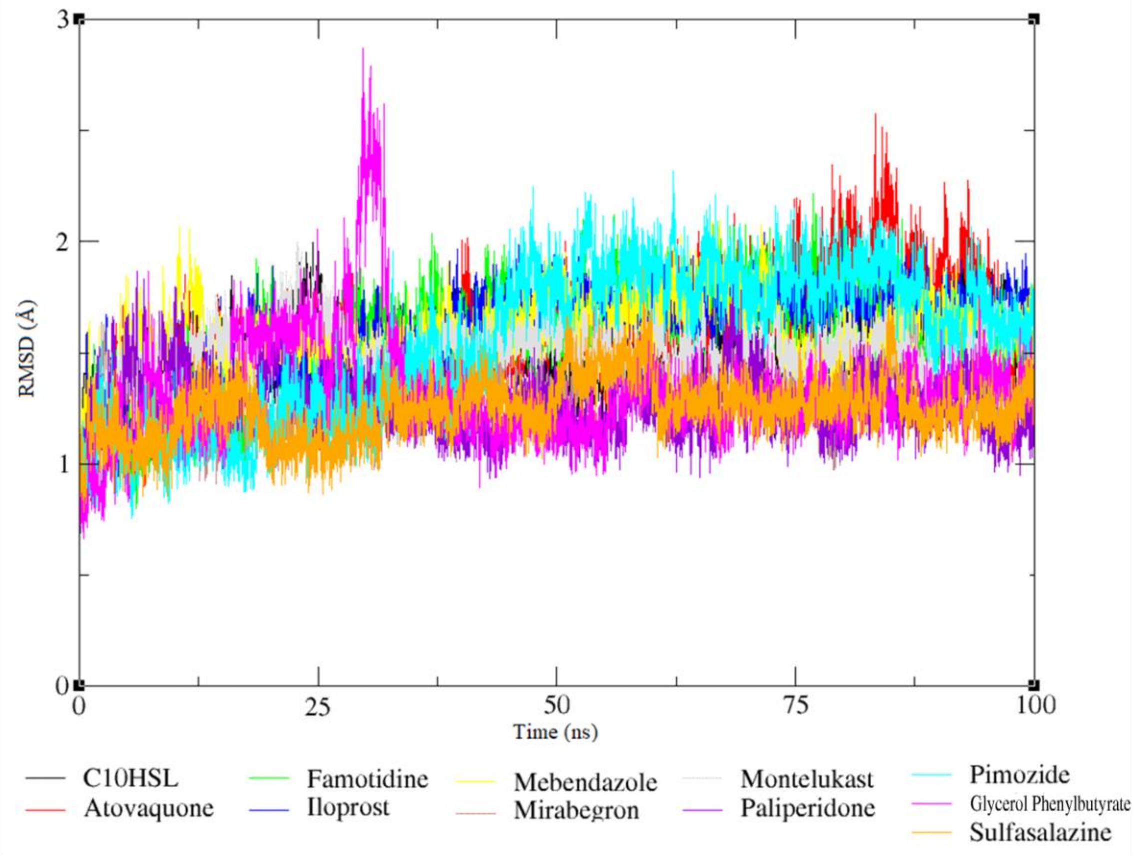


Figure 35 - Root mean square deviation plots of the Cα atoms for the CviR-ligand complexes for the selected molecules from the ZINC/FDA database.

Table 47 - Average RMSD values (Å) for the ligand, for the last 40 ns of the simulation of the CviR-ligand complexes for the selected molecules from the ZINC/FDA database.

Ligand	Average RMSD	Standard Deviation
C10HSL	0.97	0.23
Atovaquone	1.23	0.17
Famotidine	2.42	0.17
Iloprost	2.00	0.32
Mebendazole	0.75	0.22
Mirabegron	2.21	0.16
Montelukast	1.89	0.48
Paliperidone	1.07	0.29
Pimozide	2.35	0.15
Glycerol Phenylbutyrate	3.18	0.34
Sulfasalazine	1.68	0.50

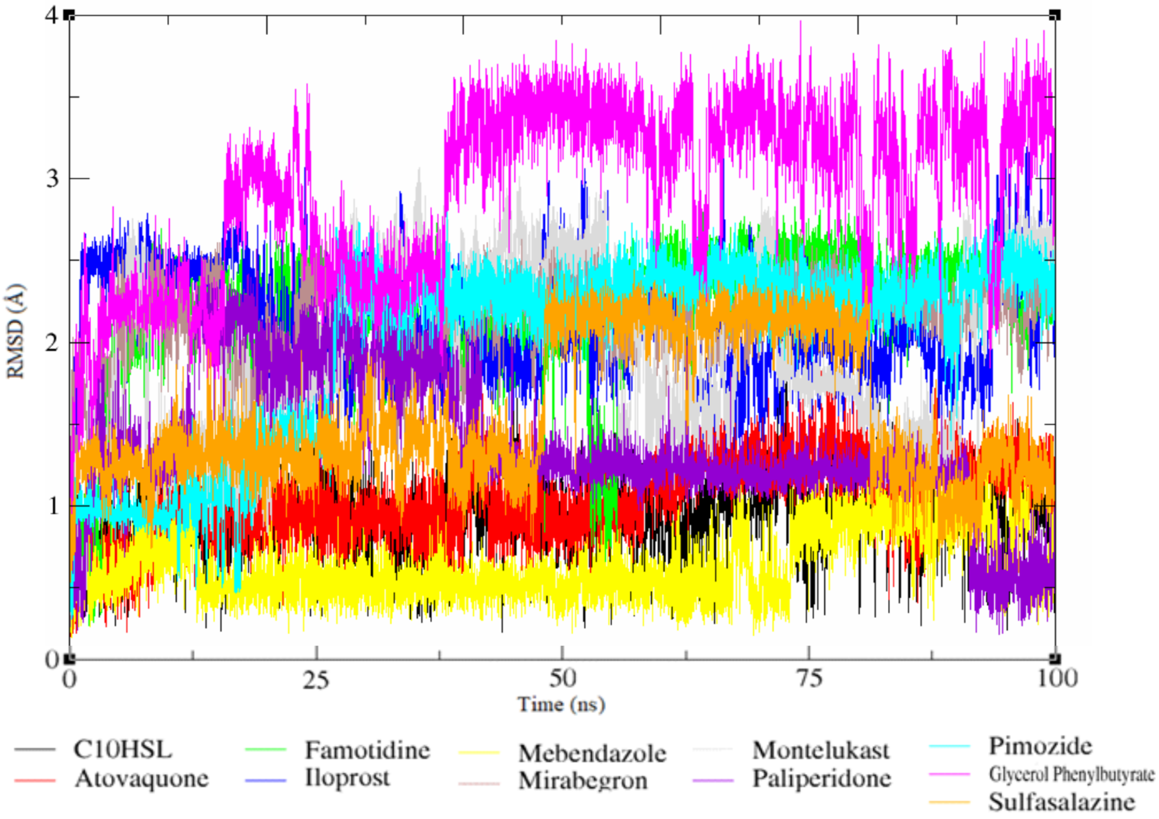


Figure 36 - Root mean square deviation plots of the ligands on the CviR-ligand complexes for the selected molecules from the ZINC/FDA database.

All complexes exhibit low RMSD values thought the simulation. In fact, all present an average RMSD value lower than 2 Å, with most never presenting a RMSD value higher than 2 Å. (see figure 34 and table 46). The exceptions were the complex with Atovaquone, Paliperidone and Pimozide. All these complexes briefly presented values higher than 2 Å but never higher than 3 Å. Most ligands (figure 35 and table 47) also display low RMSD values. However, Famotidine, Iloprost, Pimozide and Glycerol Phenylbutyrate display higher values. This indicates that the pose predicted by the docking software was not ideal, and the ligand adjusted to a more favourable pose.

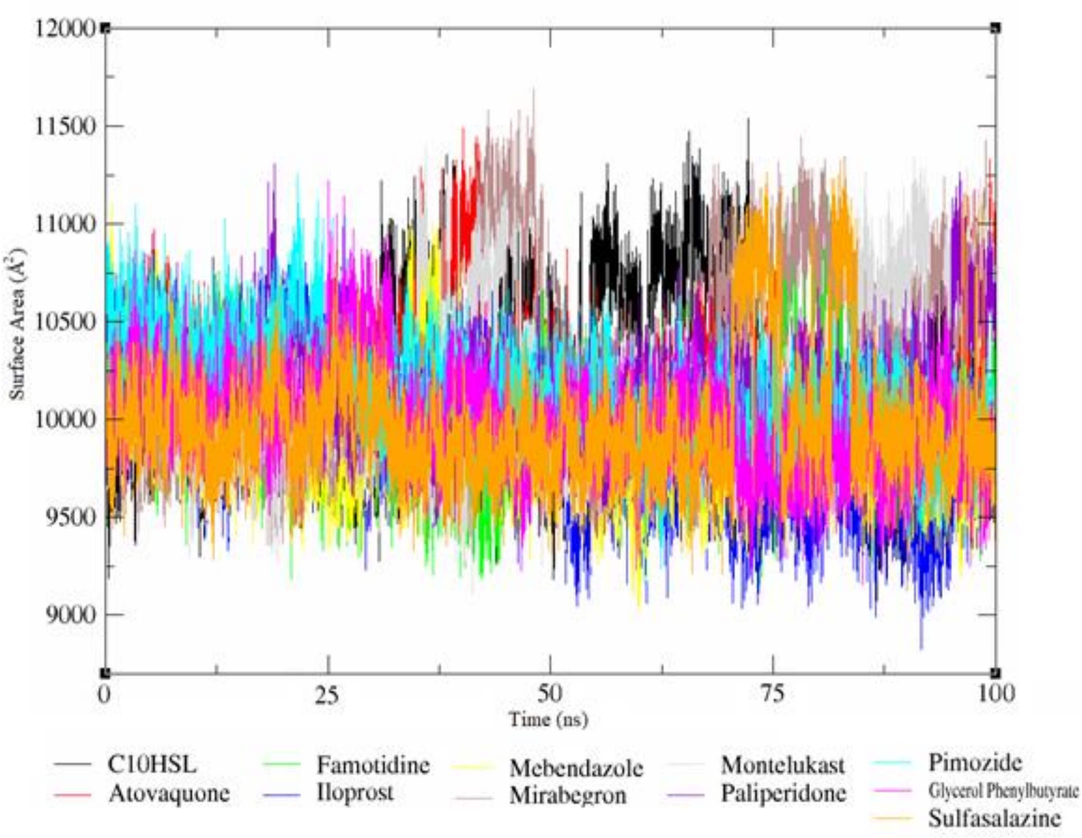


Figure 37 – Solvent accessible surface area calculation for the CviR-ligand complexes for the selected molecules from the ZINC/FDA database.

Table 48 - Average solvent accessible surface area calculation for each ligand during the last 40 ns of simulation of CviR-ligand complexes for the selected molecules from the ZINC/FDA database.

Ligand	Average SASA	Standard Deviation
C10HSL	52.20	24.40
Atovaquone	91.16	16.59
Famotidine	35.20	29.46
Iloprost	216.74	32.70
Mebendazole	138.44	21.93
Mirabegron	128.30	26.69
Montelukast	261.55	24.56
Paliperidone	211.22	22.99
Pimozide	54.89	17.83
Glycerol Phenylbutyrate	209.50	27.11
Sulfasalazine	133.09	23.04

Figure 37 displays the solvent accessible surface area for each complex along the MD simulations. If there was a sudden increase in the SASA, that would imply that the ligand would have become separated from the protein. The average SASA of each ligand is available on table 48. From these values, Famotidine and Pimozide are the ligands which display better results. Fortunately, all complexes keep a stable value along the simulation. This, together with the RMSD results, indicates that all selected ligands form stable complexes with CviR.

The final step of this study was to perform a MM/PB(GB)SA analysis for each ligand-receptor complex. For these calculations, only the last 40 ns of MD simulation were considered. The results can be seen on table 49 and a graph with the predicted Gibbs energy of association (Δ_aG) values can be seen on figure 38.

Table 49 - Predicted Gibbs energy of association (ΔaG^0) for the selected ligands from the ZINC/FDA database using MM/PBSA and MM/GBSA

MM/PB(GB)SA				
Molecule	GB	Standard Mean of Error	PB	Standard Mean of Error
C10HSL	-49.0	0.2	-23.9	0.2
Atovaquone	-38.9	0.2	-14.3	0.2
Famotidine	-29.6	0.2	-13.4	0.2
Iloprost	-48.4	0.2	-23.5	0.2
Mebendazole	-22.7	0.1	-11.7	0.1
Mirabegron	-37.7	0.2	-19.9	0.2
Montelukast	-43.0	0.3	-19.8	0.2
Paliperidone	-31.0	0.2	-14.7	0.2
Pimozide	-67.2	0.2	-24.7	0.2
Glycerol Phenybutyrate	-43.3	0.1	-24.3	0.2
Sulfasalazine	-40.1	0.2	-19.3	0.2

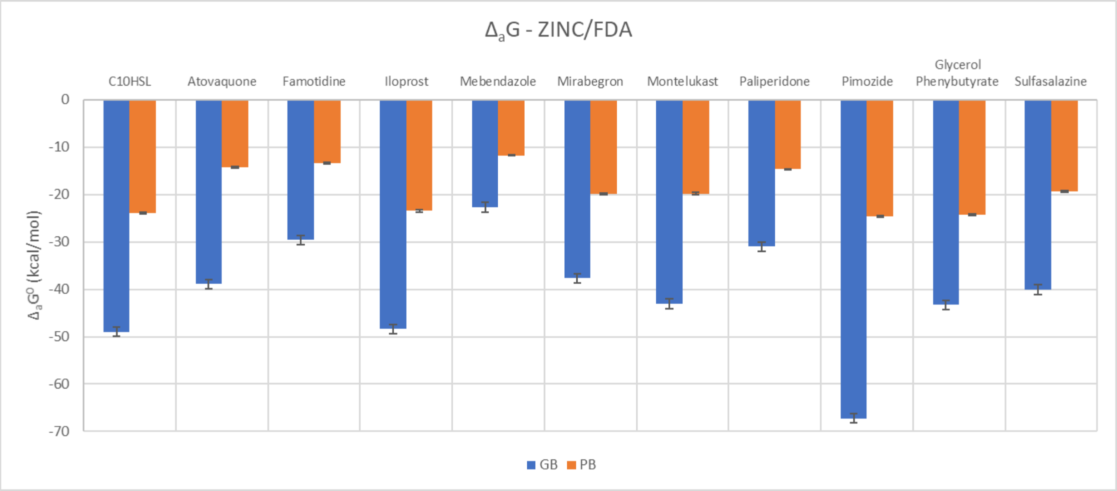


Figure 38 - Predicted Gibbs energy of association (Δ_aG) for the selected ligands from the ZINC/FDA database using MM/PBSA and MM/GBSA.

A graph with the predicted difference in Gibbs energy of association ($\Delta\Delta_aG$) using C10-HSL as reference is displayed on figure 39.

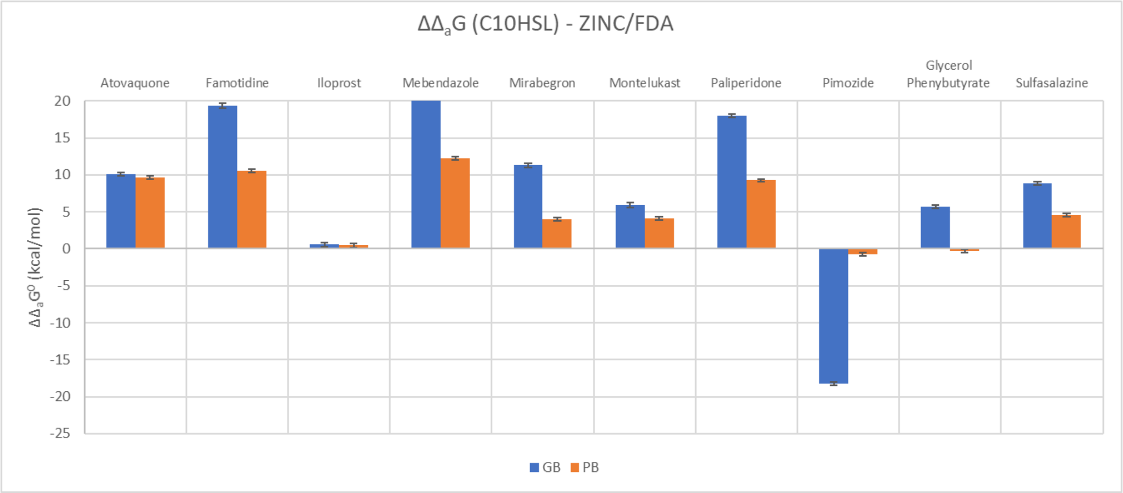


Figure 39 - Predicted difference in Gibbs energy of association ($\Delta\Delta_a G^0$) for the ZINC/FDA database, using C10-HSL as reference, calculated using MM/PBSA and MM/GBSA.

The MM/PBSA and MM/GBSA free energies are predictions which should not be compared to experimental results, bearing in mind that the entropy is not being considered in the calculation. However, it is beneficial to compare the predicted Gibbs energy of association of different ligands bound to the same protein to have relative binding affinities.

Analysing the difference in Gibbs energy of association, it becomes clear that most ligands showcase a lower affinity towards CviR than C10-HSL. The only ligand which displays larger affinity towards CviR using both MM/PBSA and MM/GBSA is Pimozide. Glycerol Phenylbutyrate shows good results using MM/PBSA but presents lower affinity towards CviR using MM/GBSA. Another notable result is the performance of Iloprost which, in both methods, displayed similar performance to C10-HSL.

To further analyse the affinity between the ligands and the receptor, the overall Gibbs energy of association was decomposed into the contribution of each residue. The individual energy of association of residues which, in general, have a bigger impact in the predicted Gibbs energy of association are represented on figure 40 and 41.

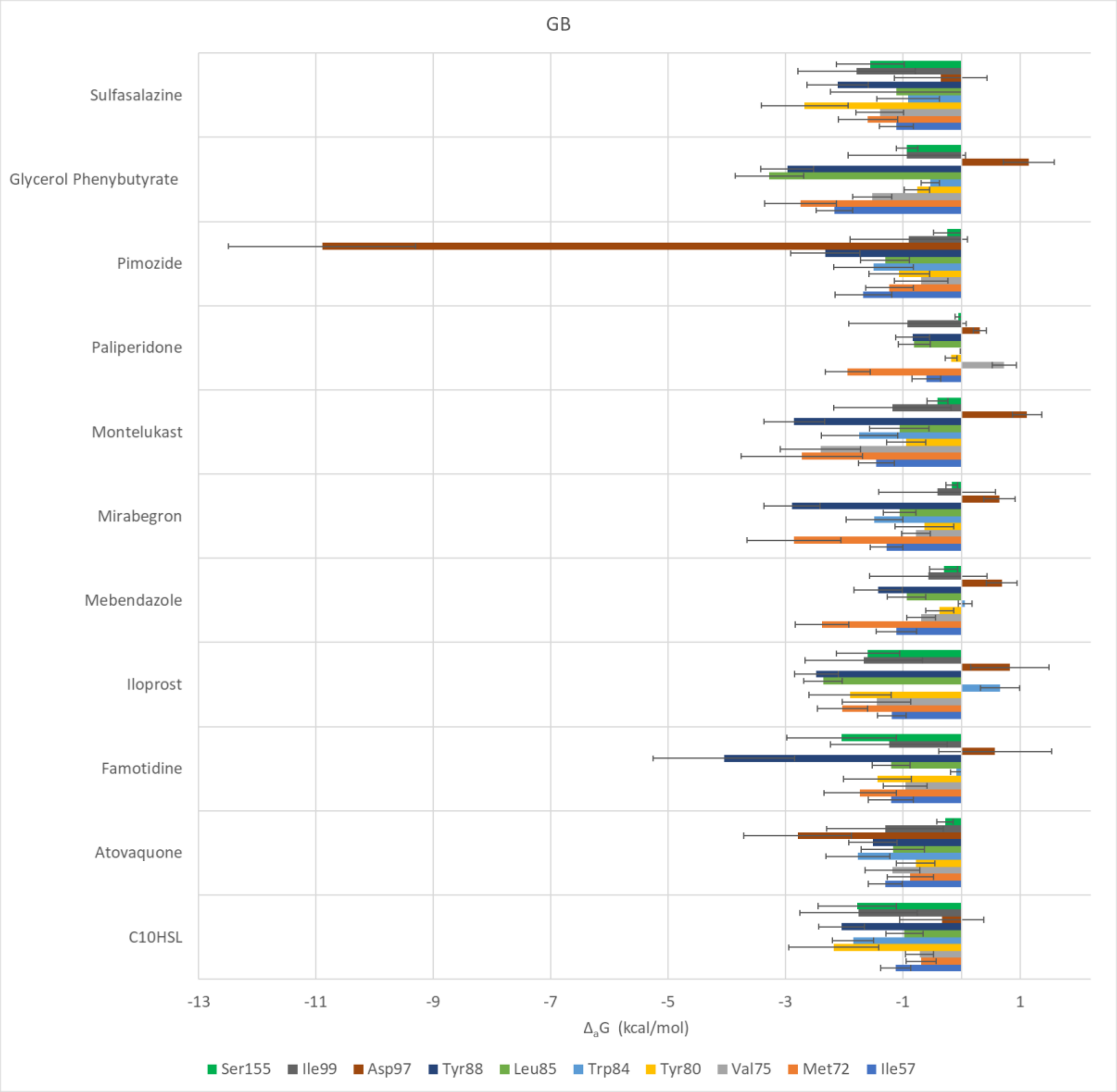


Figure 40 – Decomposition of the Gibbs energy of association (Δ_aG), calculated using MM/GBSA, for the selected ligands of the ZINC/FDA database.

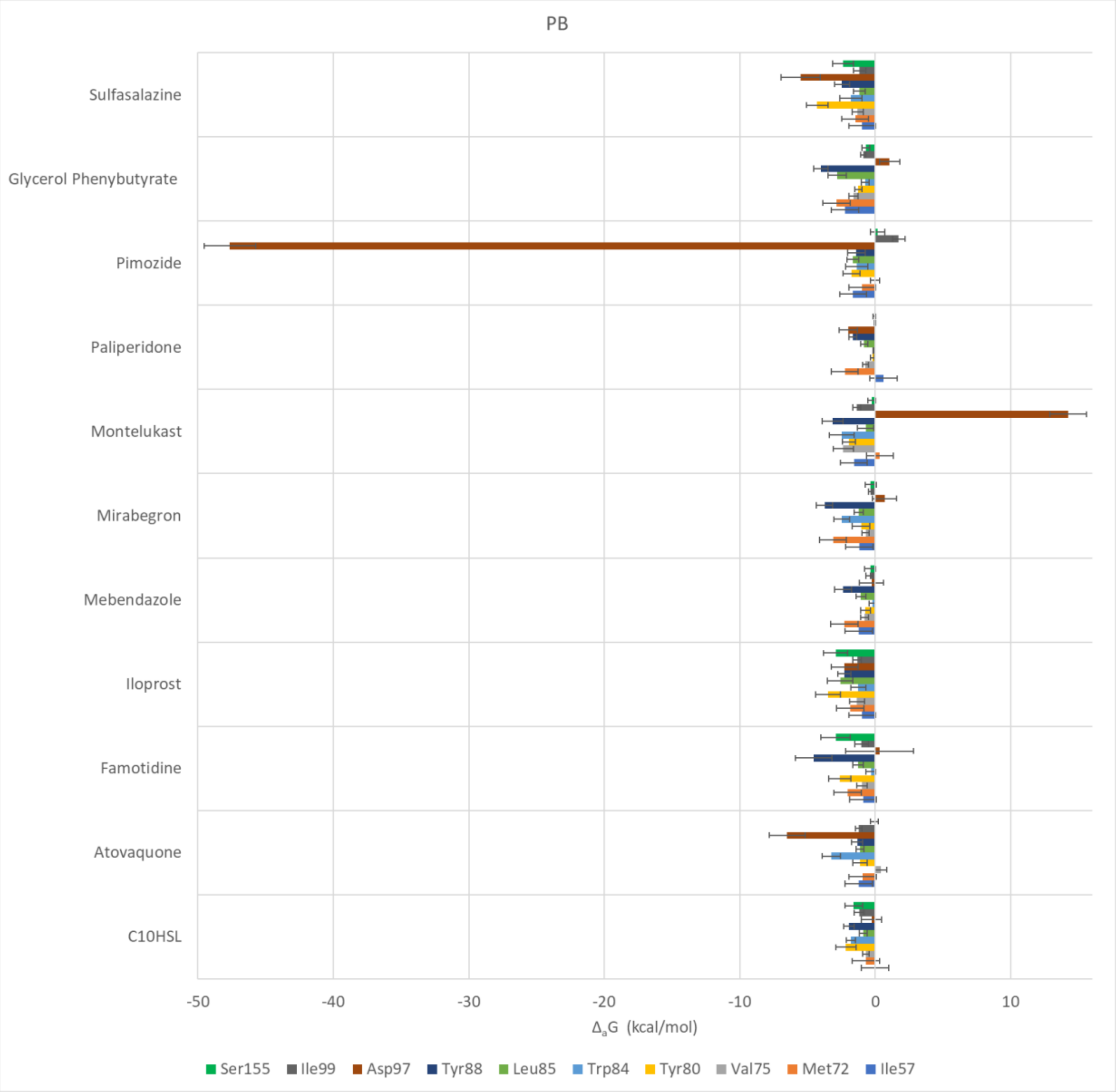


Figure 41 - Decomposition of the Gibbs energy of association ($\Delta_a G$), calculated using MM/PBSA, for the selected ligands of the ZINC/FDA database.

On C10-HSL, using both methods, the amino acids that have a bigger impact are Tyr80, Trp84, Tyr88, Ile99 and Ser155. When comparing with the already known interactions map, these results show great similarity. The only difference is the contribution of Asp97. While this residue is considered to have an important role on the binding of ligand to CviR, these calculations result in a lower value for Asp97. In fact, this residue shows a negative contribution in nearly all calculations using MM/GBSA with the only exception being with Pimozide where it is the major contribution to its affinity towards CviR. On the MM/PBSA calculations Asp97 has a more positive contribution on multiple ligands, with its contribution to the affinity of Pimozide being even more significative. On Iloprost, the amino acids which show a higher contribution to the Gibbs energy of association are Met72, Tyr80, Leu85 and Tyr88 using GB, and Met72, Tyr80, Leu85, Tyr88 and Asp 97 using PB. In the case of Pimozide, besides Asp97, only Tyr88, using the GB method, has a significant contribution compared to the other highlighted amino acids. Lastly, on the MM/PBSA calculations with Glycerol Phenylbutyrate, the amino acids with the biggest impact are Met72, Leu85 and Tyr88.

4.5.2.2 Chemotheca

As previously mentioned, the 10 molecules from Chemotheca that were selected for further experiments were CMLDID2574, CMLDID5450, CMLDID17434, CMLDID18049, CMLDID23812, CMLDID35542, CMLDID38590, CMLDID40723, CMLDID50121 and CMLDID60399. As before, 100 ns of MD simulation was performed for each molecule in complex with 3QP6 and the results were compared to the results obtained with C10-HSL. To evaluate the structural stability of the complex, RMSD calculations were performed for each complex and ligand (see figure 42, figure 43, table 50 and table 51).

Table 50 - Average RMSD values (Å) for the last 40ns of simulation of the CviR-ligand complexes for the selected molecules from the Chemotheca database

Ligand	Average RMSD	Standard Deviation
C10HSL	0.97	0.23
CMLDID17434	1.50	0.12
CMLDID18049	1.31	0.13
CMLDID23812	1.39	0.16
CMLDID2574	1.32	0.14
CMLDID35542	1.34	0.10
CMLDID38590	1.41	0.13
CMLDID40723	1.42	0.13
CMLDID50121	1.37	0.16
CMLDID5450	1.46	0.09
CMLDID60399	1.29	0.07

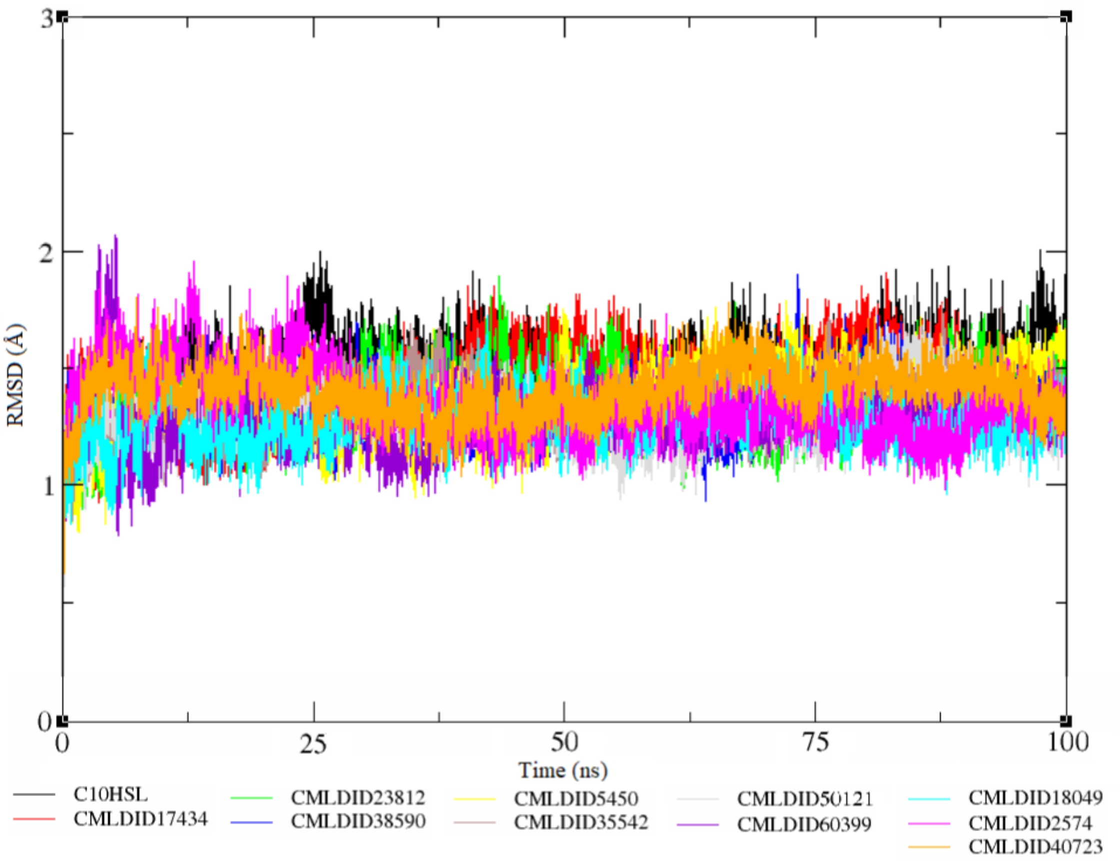


Figure 42 - Root mean square deviation plots of the Cα atoms for the CviR-ligand complexes for the selected molecules from the Chemotheca database.

Table 51- Average RMSD values (Å) for the ligand, for the last 40 ns of the simulation of the CviR-ligand complexes for the selected molecules from the Chemotheca database

Ligand	Average RMSD	Standard Deviation
C10HSL	0.97	0.23
CMLDID17434	2.64	0.43
CMLDID18049	2.15	0.14
CMLDID23812	2.50	0.43
CMLDID2574	1.10	0.30
CMLDID35542	2.07	0.24
CMLDID38590	3.19	0.29
CMLDID40723	1.56	0.40
CMLDID50121	2.42	0.41
CMLDID5450	2.91	0.43
CMLDID60399	1.26	0.17

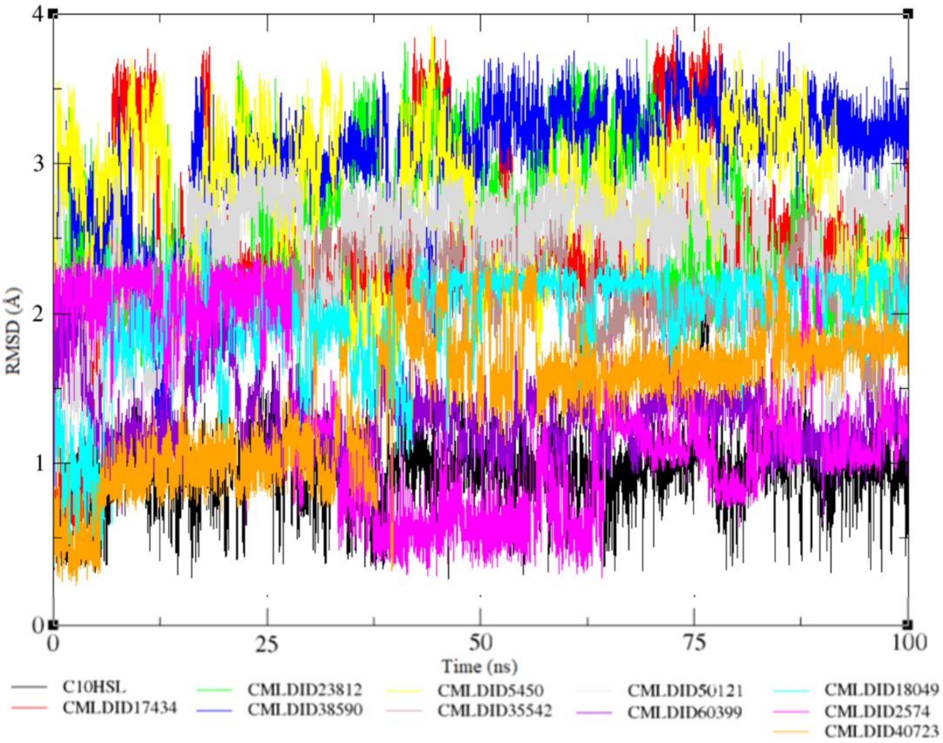


Figure 43 - Root mean square deviation plots of the ligands of the CviR-ligand complexes for the selected molecules from the Chemotheca database

All protein-ligand complexes display low RMSD values thought the simulation, with all average RMSD values being below 2 Å. Some ligands also display low RMSD values, however, multiple ligands have higher RMSD values. As before, this indicates that the pose predicted by the docking software was not ideal, and the ligand adjusted to a more favourable pose.

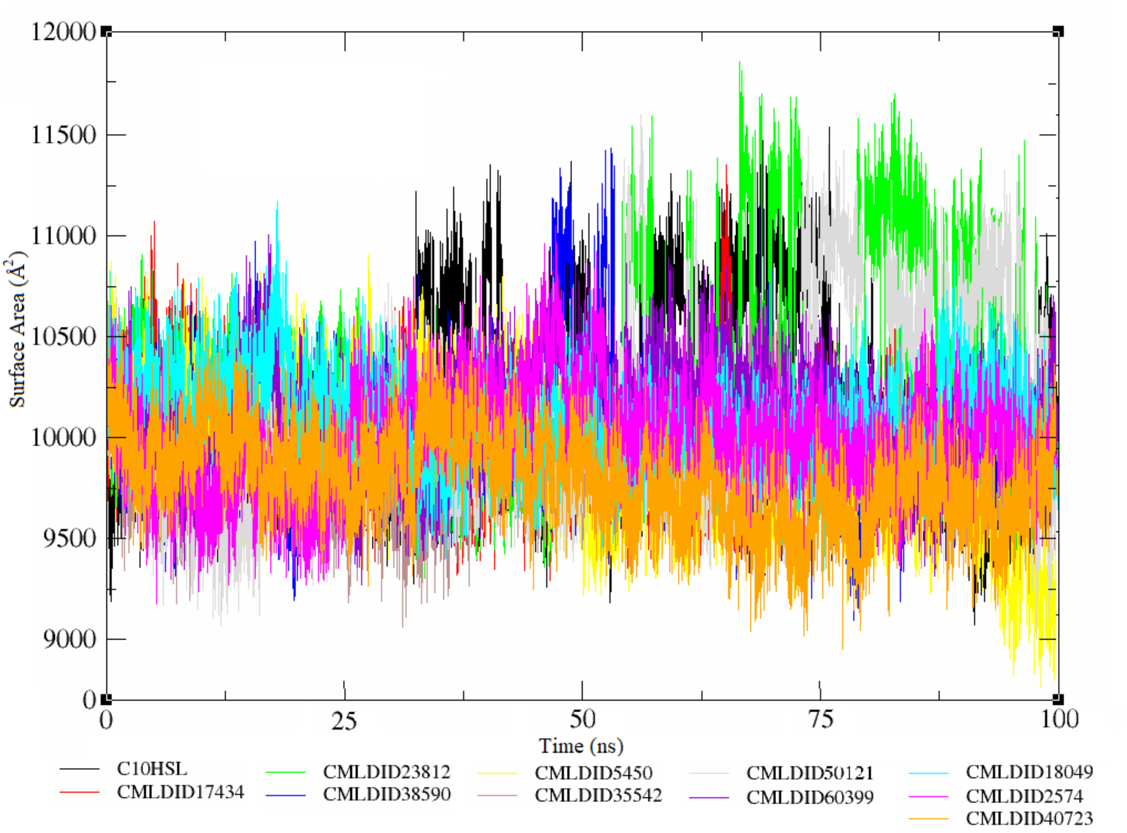


Figure 44 – Solvent accessible surface area calculation for the CviR-ligand complexes for the selected molecules from the Chemotheca database

Table 52 – Average solvent accessible surface area calculation for each ligand during the last 40 ns of simulation of CviR-ligand complexes for the selected molecules from the Chemotheca database

Ligand	Average SASA	Standard Deviation
C10HSL	52.20	24.39
CMLDID17434	166.97	22.19
CMLDID18049	191.94	69.17
CMLDID23812	172.38	48.51
CMLDID2574	168.89	26.81
CMLDID35542	96.61	25.48
CMLDID38590	194.71	27.78
CMLDID40723	119.11	33.02
CMLDID50121	109.18	33.41
CMLDID5450	185.42	48.40
CMLDID60399	77.15	14.32

Figure 44 displays the SASA for each protein-ligand complex along the MD simulations. On table 52, the average SASA for each ligand can be seen. CMLDID18049 is the ligand with the best result. There was no sudden change in the accessible surface area in none of the complexes. This is in agreement with the RMSD results, showing that all complexes were stable for the length of the simulation.

As before, MM/PB(GB)SA calculations were performed for the last 40 ns of MD simulation for each ligand-receptor complex. The results can be seen on table 53, and a graph with the predicted Gibbs energy of association ($\Delta_a G$) values can be seen on figure 45.

Table 53 - Predicted Gibbs energy of association (ΔaG) for the selected ligands from the Chemotheca database using MM/PBSA and MM/GBSA.

MM/PB(GB)SA				
Molecule	GB	Standard Mean of Error	PB	Standard Mean of Error
C10HSL	-49.0	0.2	-23.9	0.2
CMLDID17434	-50.0	0.2	-24.5	0.2
CMLDID18049	-27.8	0.2	-13.8	0.2
CMLDID23812	-48.0	0.2	-24.1	0.2
CMLDID2574	-25.4	0.1	-12.3	0.1
CMLDID35542	-44.4	0.2	-22.3	0.2
CMLDID38590	-47.3	0.2	-23.4	0.2
CMLDID40723	-34.1	0.2	-17.6	0.2
CMLDID50121	-45.4	0.2	-20.6	0.2
CMLDID5450	-48.4	0.2	-22.9	0.2
CMLDID60399	-53.0	0.2	-27.8	0.2

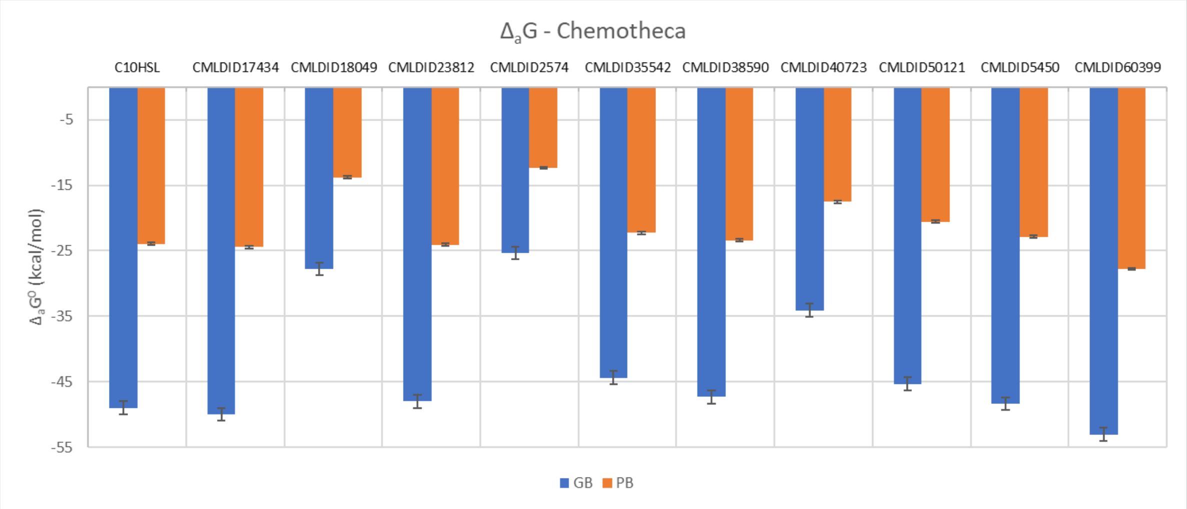


Figure 45 - Predicted Gibbs energy of association (ΔaG) for the selected ligands from the Chemotheca database using MM/PBSA and MM/GBSA.

The predicted difference in Gibbs energy of association ($\Delta\Delta_a G$) using C10-HSL as reference is displayed on figure 46.

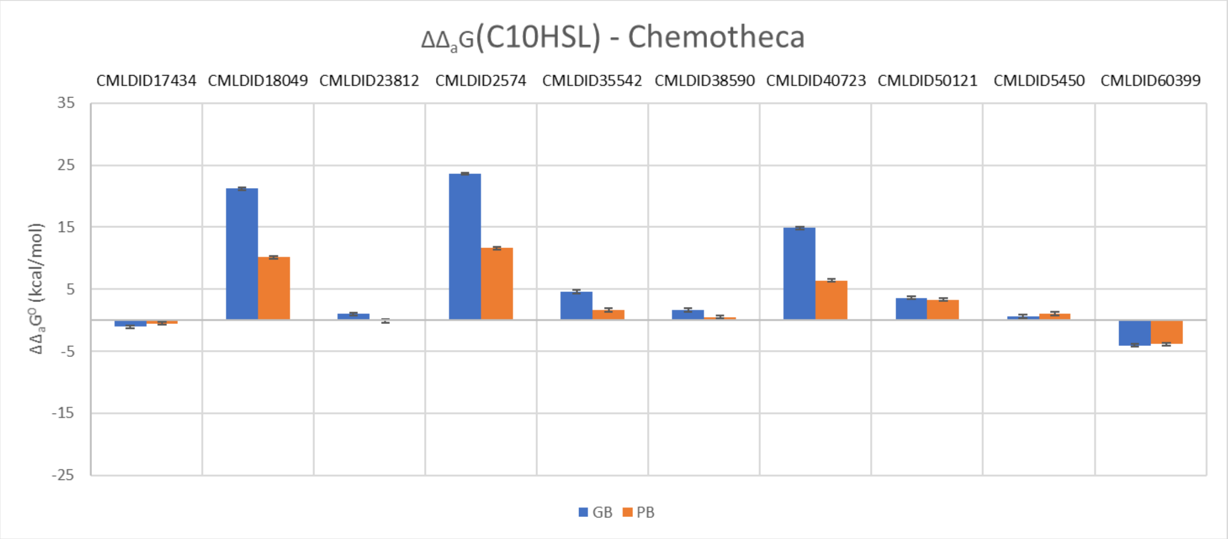


Figure 46 - Predicted difference in Gibbs energy of association ($\Delta\Delta_a G$) for the Chemotheca database, using C10-HSL as reference, calculated using MM/PBSA and MM/GBSA.

In Figure 45 it can be seen that the ligands that display higher affinity towards CviR than C10-HSL are CMLDID17434 and CMLDID60399 with both methods, and CMLDID 23812 using MM/PBSA.

As previously done, the overall Gibbs energy of association was decomposed into the contribution of each residue. The individual energy of association of residues which have a bigger impact in the predicted Gibbs energy of association are represented on figure 47 and 48.

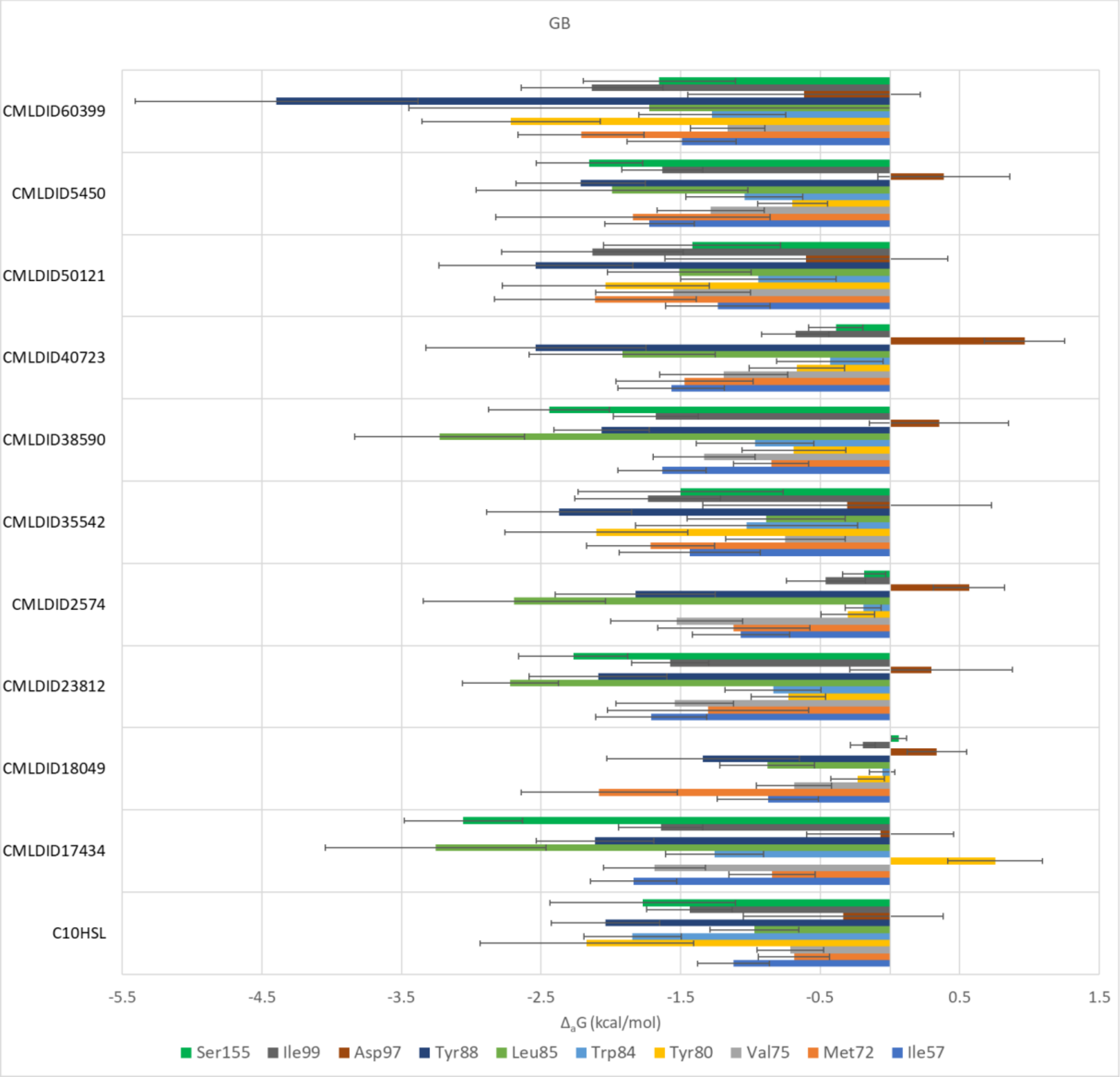


Figure 47 - Decomposition of the Gibbs energy of association ($\Delta_a G$), calculated using MM/GBSA, for the selected ligands of the Chemotheca database.

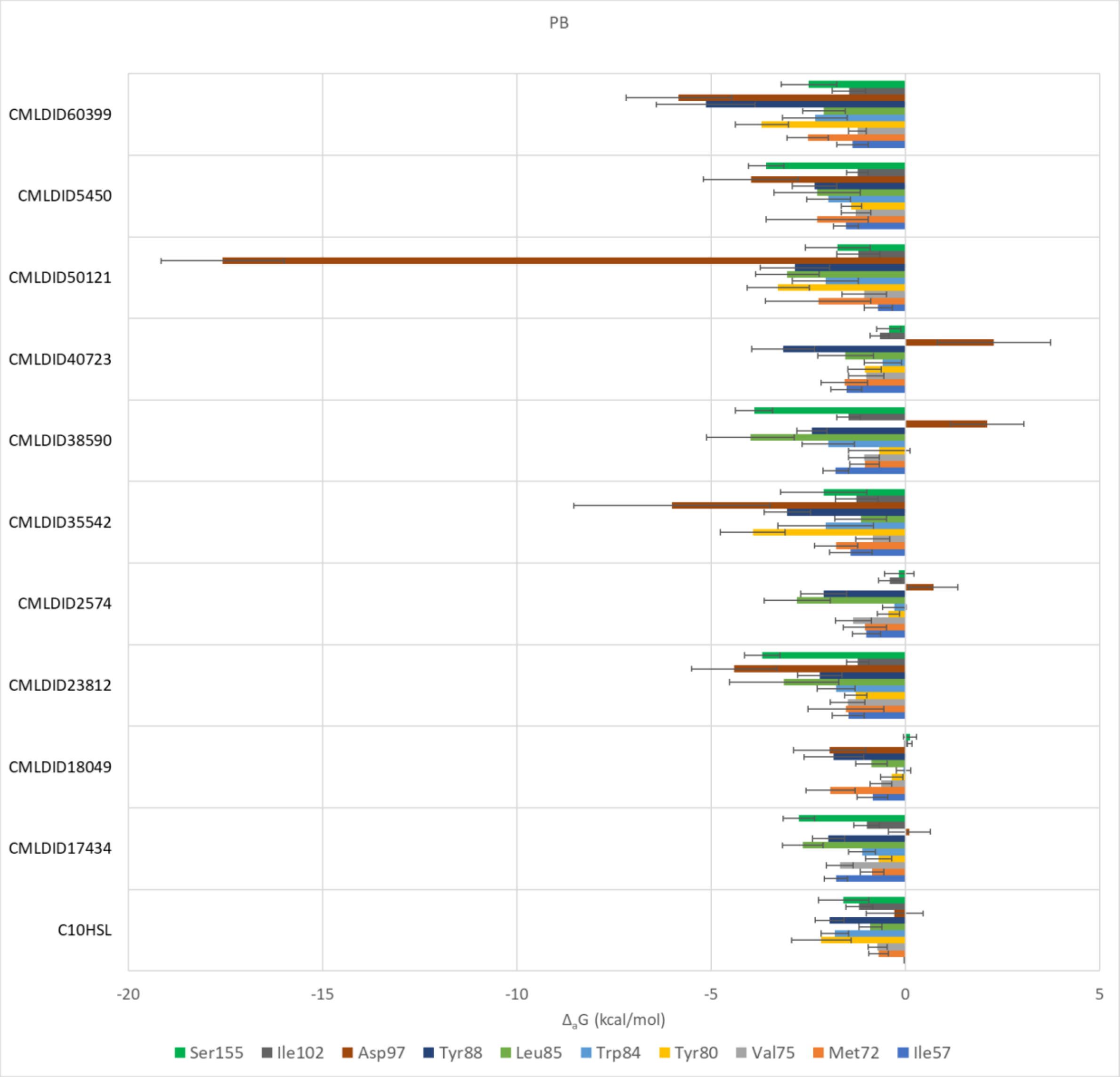


Figure 48 - Decomposition of the Gibbs energy of association (Δ_aG), calculated using MM/PBSA, for the selected ligands of the Chemotheca database.

For CMLDID17434, the residues with the highest contribution to the overall result using both methods are Leu85, Tyr88 and Ser155. Although these are the amino acids that contribute the most to CMLDID17434's affinity towards CviR, most residues have a very positive contribution to this ligand's affinity. The only exceptions are Tyr80 and Asp97 in the MM/GBSA calculations and Asp97 in MM/PBSA. As was the case with the ligands from the ZINC/FDA database, in these calculations Asp97 also shows a curious behaviour. In the MM/GBSA calculations this residue shows negative or little influence on the final results of most ligands. In the MM/PBSA calculations it has a more positive contribution for the affinity of most ligands, but it still has a negative contribution in some ligands. Adding to this, similarly to what was observed during the previous section, Asp97 again displays a much higher value than what is observed for all the residues. This is observed on the MM/PBSA calculations for CMLDID50121. For CMLDID2381, using the

MM/PBSA method, the residues with the most impact in the affinity towards CviR are Leu85, Tyr88, Asp97 and Ser155. Lastly, for CMLDID60399, the amino acids with the greater impact are Met72, Tyr80, Tyr88, for both methods. In the MM/PBSA calculations, Asp97 also displays a high contribution to the affinity of the ligand towards CviR.

4.5.3 Conclusions

The molecular dynamics simulations and the MM/PB(GB)SA calculations performed in this section resulted in six molecules with higher or comparable binding affinities to C10-HSL. These molecules were Pimozide (figure 49), Glycerol Phenylbutyrate (figure 50) and Iloprost (figure 51) from the ZINC/FDA database and CMLDID17434 (figure 52), CMLDID23812 (figure 53) and CMLDID60399 (figure 54) from the Chemotheca database. Overall, the amino acids which contribute the most to a higher affinity of the ligand towards CviR, across all ligands, are Met72, Tyr80, Leu85 and Tyr88 and Ser155.

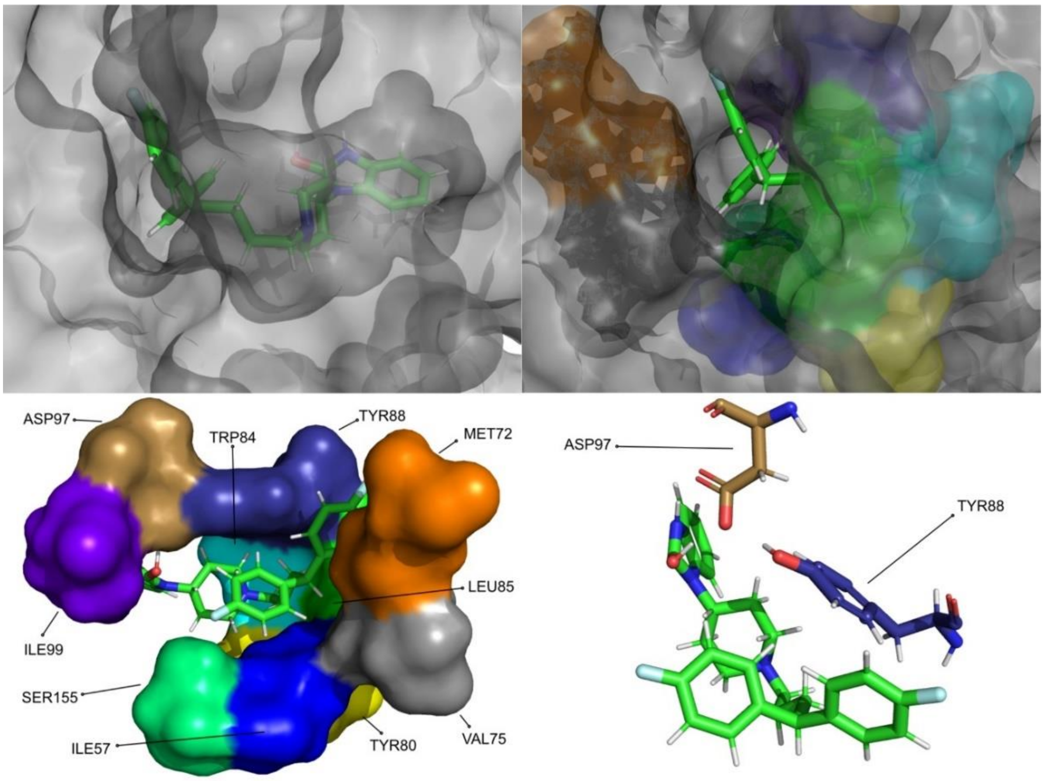


Figure 49 – Pimozide in complex with CviR. Top left picture represents the ligand in licorice and the protein in surface. Top right and bottom left pictures feature, in surface, the amino acids residues which, overall, have a bigger impact in the predicted affinity. Bottom right picture represents, in licorice. the amino acids with the biggest contribution to the affinity of this ligand.

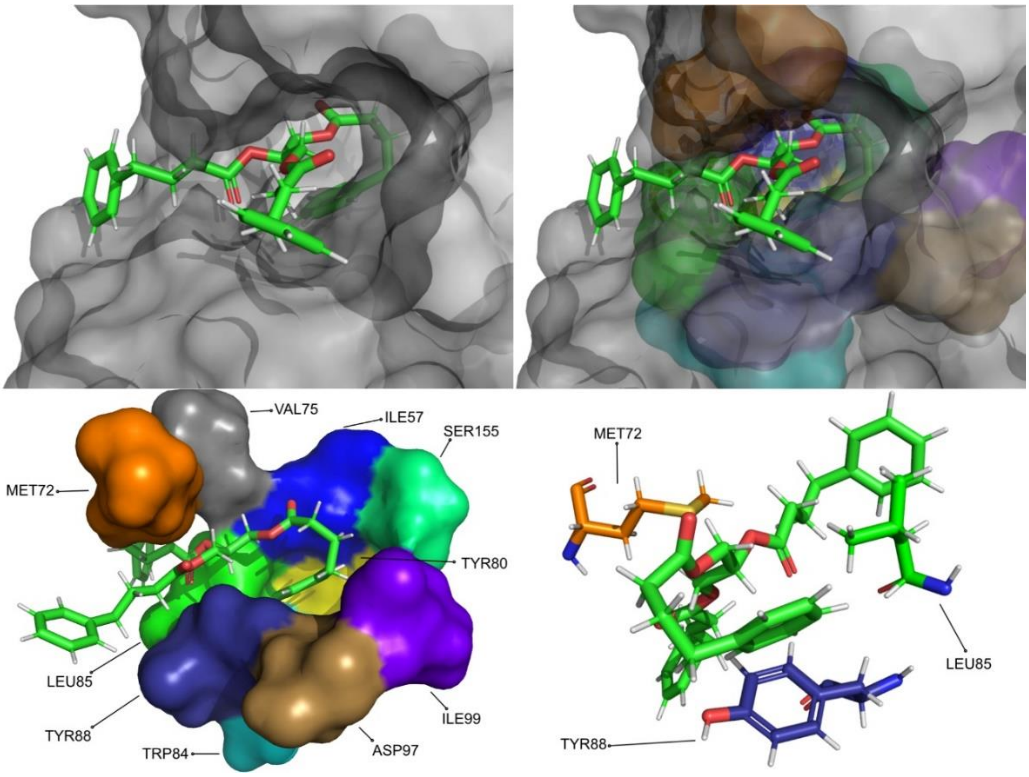


Figure 50 - Glycerol Phenylbutyrate in complex with CviR. Top left picture represents the ligand in licorice and the protein in surface. Top right and bottom left pictures feature, in surface, the amino acids residues which, overall, have a bigger impact in the predicted affinity. Bottom right picture represents, in licorice. the amino acids with the biggest contribution to the affinity of this ligand

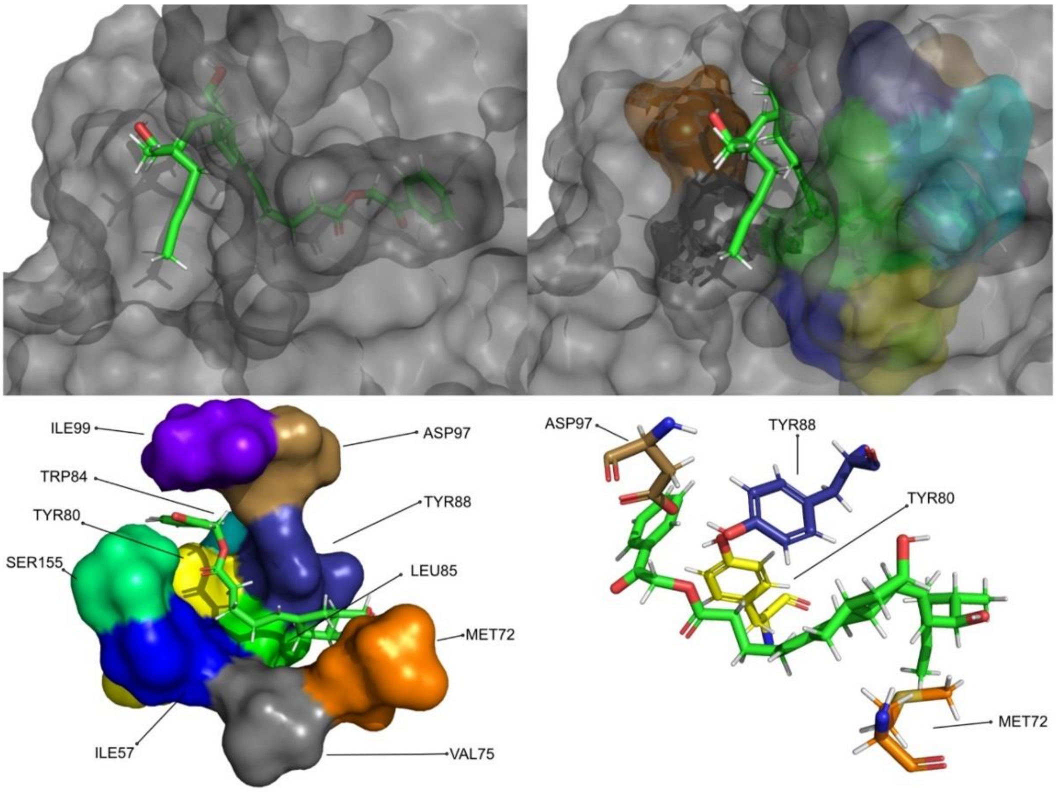


Figure 51 - Iloprost in complex with CviR. Top left picture represents the ligand in licorice and the protein in surface. Top right and bottom left pictures feature, in surface, the amino acids residues which, overall, have a bigger impact in the predicted affinity. Bottom right picture represents, in licorice. the amino acids with the biggest contribution to the affinity of this ligand.

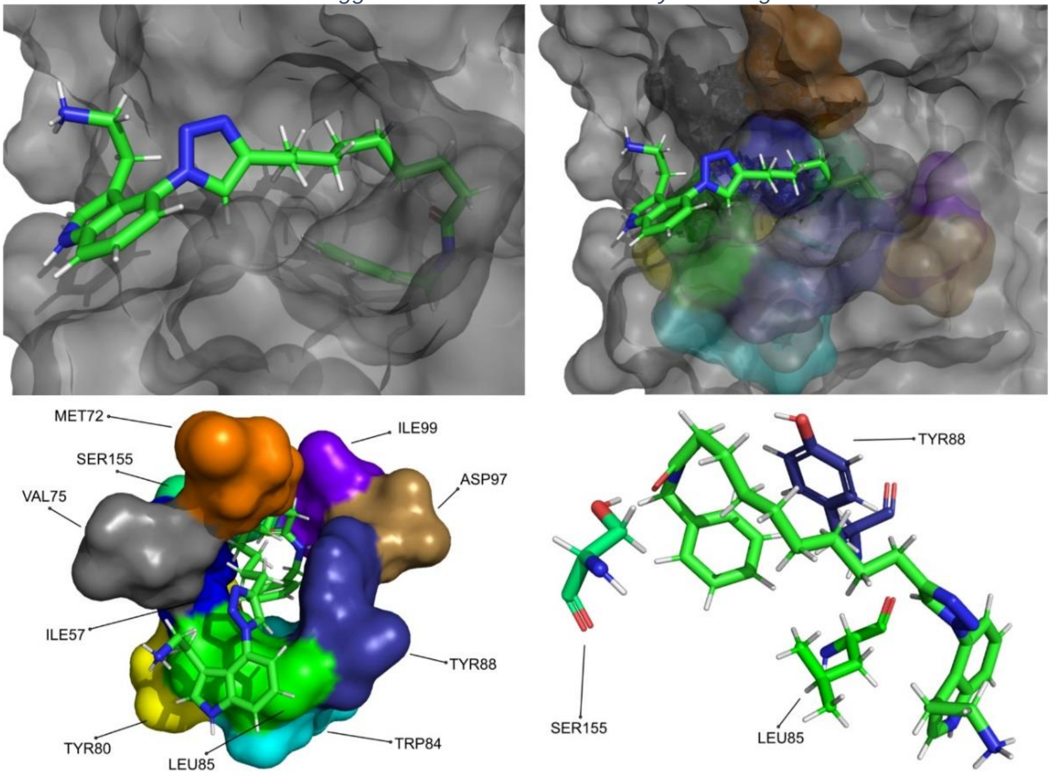


Figure 52 - CMLDID17434 in complex with CviR. Top left picture represents the ligand in licorice and the protein in surface. Top right and bottom left pictures feature, in surface, the amino acids residues which, overall, have a bigger impact in the predicted affinity. Bottom right picture represents, in licorice. the amino acids with the biggest contribution to the affinity of this ligand

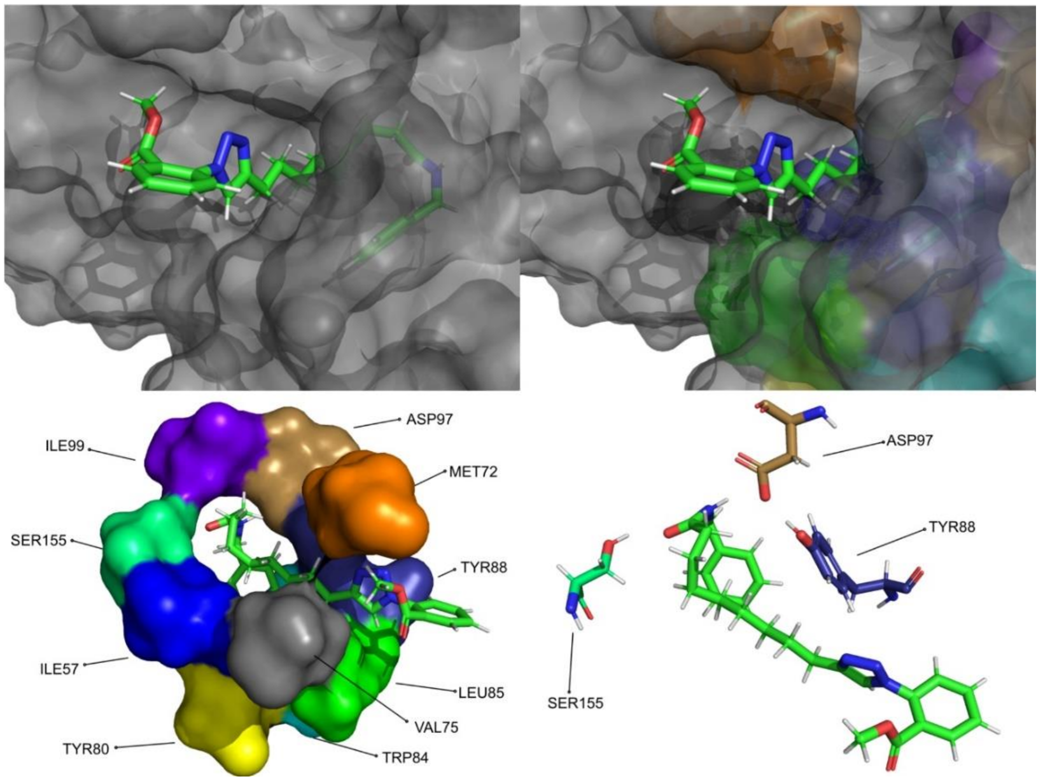


Figure 53 - CMLDID23812 in complex with CviR. Top left picture represents the ligand in licorice and the protein in surface. Top right and bottom left pictures feature, in surface, the amino acids residues which, overall, have a bigger impact in the predicted affinity. Bottom right picture represents, in licorice. the amino acids with the biggest contribution to the affinity of this ligand

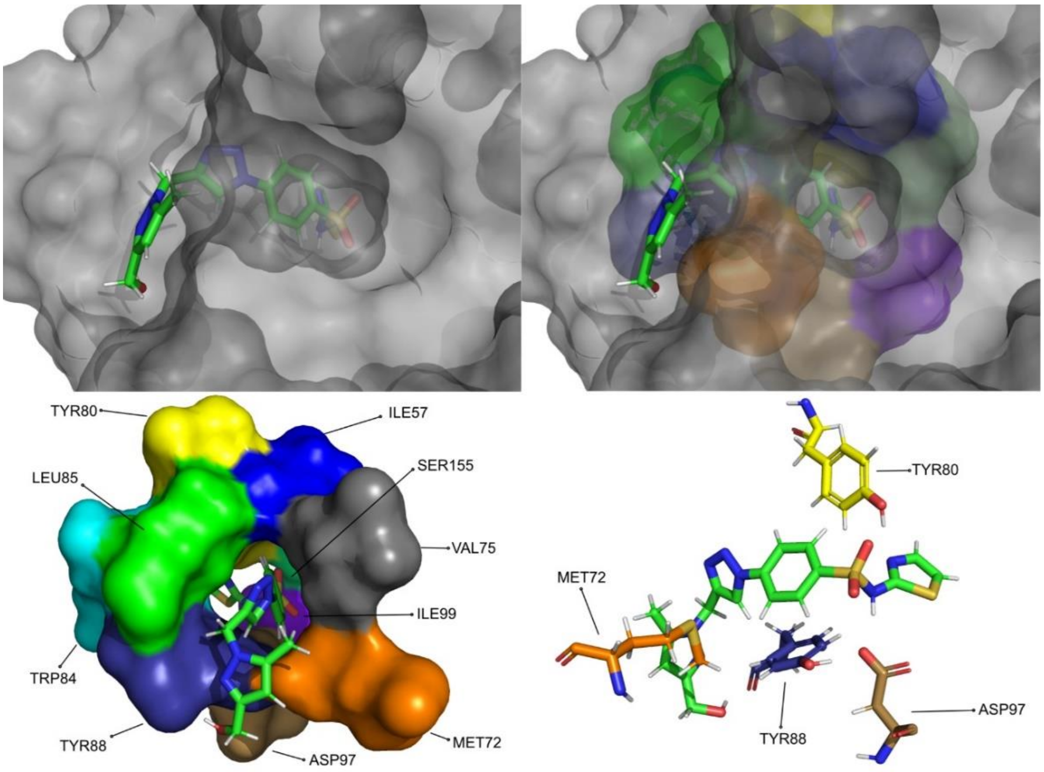


Figure 54 - CMLDID60399 in complex with CviR. Top left picture represents the ligand in licorice and the protein in surface. Top right and bottom left pictures feature, in surface, the amino acids residues which, overall, have a bigger impact in the predicted affinity. Bottom right picture represents, in licorice. the amino acids with the biggest contribution to the affinity of this ligand

5. Conclusion

The aim of this work was to use Computer Assisted Drug Design to model promising molecules to block quorum sensing and therefore prevent biofilm formation. This was achieved by optimizing molecular docking and virtual screening protocols focused on the quorum sensing receptor from *Chromobacterium violaceum,* CviR. The final result a new selection of promising compounds which were than analysed using molecular dynamics simulations and MM/PB(GB)SA calculations.

Because microorganisms embedded in biofilms have several advantages, infections associated with biofilms have been accepted as a significant danger to our society. The recalcitrance of these structures towards existing antimicrobial approaches made necessary the discovery of novel methods to inhibit their mechanisms of formation. Inhibiting the formation of biofilms by disrupting quorum sensing is the most promising strategy.

The optimized molecular docking and virtual screening protocols were applied to two databases, a database comprising of FDA approved compounds, obtained from ZINC, and the Mu.Ta.Lig Chemotheca. These virtual screening procedures resulted in a list of compounds which can be further tested, either in a computational or experimental setting.

20 of the most promising compounds that resulted from the virtual screening procedures were than further analysed using molecular dynamics simulations and MM/PB(GB)SA calculations. These procedures predicted that six molecules of the initial 20 have similar or better affinity towards CviR than the reference ligand. From the ZINC/FDA database, the best results were obtained with Pimozide, Glycerol Phenylbutyrate and Iloprost. From the Chemotheca database, CMLDID17434, CMLDID 23812 and CMLDID60399 had the most promising results. These results, together with compounds which can be found in the future, are valuable information for an eventual experimental screening.

In the future, the optimized protocol can be applied to other databases, such as the ZINC lead-like database, consisting of over 4.6 million molecules. This will hopefully result in additional promising compounds.

In brief, this work reports the development of an optimized CADD protocol for the development of new quorum sensing inhibitors. Through the usage of multiple computational techniques, it was possible to obtain a list of compounds to be validated experimentally, which will hopefully result in the discovery of new drugs against biofilm formation.

6. Bibliographic References

(1) Vert, M.; Doi, Y.; Hellwich, K. H.; Hess, M.; Hodge, P.; Kubisa, P.; Rinaudo, M.; Schué, F. Terminology for Biorelated Polymers and Applications (IUPAC Recommendations 2012). *Pure Appl. Chem.* **2012**, *84* (2), 377–410. https://doi.org/10.1351/pac-rec-10-12-04.

(2) Konopka, A. What Is Microbial Community Ecology. *ISME J.* **2009**, *3* (11), 1223–1230. https://doi.org/10.1038/ismej.2009.88.

(3) Kolter, R.; Greenberg, E. P. Microbial Sciences: The Superficial Life of Microbes. *Nature* **2006**, *441* (7091), 300–302. https://doi.org/10.1038/441300a.

(4) Hall, C. W.; Mah, T. F. Molecular Mechanisms of Biofilm-Based Antibiotic Resistance and Tolerance in Pathogenic Bacteria. *FEMS Microbiol. Rev.* **2017**, *41* (3), 276–301. https://doi.org/10.1093/femsre/fux010.

(5) Kouzel, N.; Oldewurtel, E. R.; Maier, B. Gene Transfer Efficiency in Gonococcal Biofilms: Role of Biofilm Age, Architecture, and Pilin Antigenic Variation. *J. Bacteriol.* **2015**, *197* (14), 2422–2431. https://doi.org/10.1128/JB.00171-15.

(6) Smolentseva, O.; Gusarov, I.; Gautier, L.; Shamovsky, I.; Defrancesco, A. S.; Losick, R.; Nudler, E. Mechanism of Biofilm-Mediated Stress Resistance and Lifespan Extension in C. Elegans. *Sci. Rep.* **2017**, *7* (1), 1–16. https://doi.org/10.1038/s41598-017-07222-8.

(7) Saxena, P.; Joshi, Y.; Rawat, K.; Bisht, R. Biofilms: Architecture, Resistance, Quorum Sensing and Control Mechanisms. *Indian J. Microbiol.* **2019**, *59* (1), 3–12. https://doi.org/10.1007/s12088-018-0757-6.

(8) Kostakioti, M.; Hadjifrangiskou, M.; Hultgren, S. J. Bacterial Biofilms: Development, Dispersal, and Therapeutic Strategies in the Dawn of the Postantibiotic Era. *Cold Spring Harb. Perspect. Med.* **2013**, *3* (4). https://doi.org/10.1101/cshperspect.a010306.

(9) Jamal, M.; Ahmad, W.; Andleeb, S.; Jalil, F.; Imran, M.; Nawaz, M. A.; Hussain, T.; Ali, M.; Rafiq, M.; Kamil, M. A. Bacterial Biofilm and Associated Infections. *J. Chinese Med. Assoc.* **2018**, *81* (1), 7–11. https://doi.org/10.1016/j.jcma.2017.07.012.

(10) Otto, M. Staphylococcal Infections: Mechanisms of Biofilm Maturation and Detachment as Critical Determinants of Pathogenicity. *Annu. Rev. Med.* **2013**,

64 (1), 175–188. https://doi.org/10.1146/annurev-med-042711-140023.

(11) Flemming, H. C.; Wingender, J. The Biofilm Matrix. *Nat. Rev. Microbiol.* **2010**, *8* (9), 623–633. https://doi.org/10.1038/nrmicro2415.

(12) Flemming H-C, Wingender J, Griegbe, M. C. Physico-Chemical Properties of Biofilms. In *Biofilms: recent advances in their study and control*; Evans, L., Ed.; Amsterdam: Harwood Academic Publishers, 2000; pp 19–34.

(13) Sutherland, I. W. Biofilm Exopolysaccharides: A Strong and Sticky Framework. *Microbiology* **2001**, *147* (1), 3–9. https://doi.org/10.1099/00221287-147-1-3.

(14) Donlan, R. M. Biofilms : Microbial Life on Surfaces. *Emerg. Infect. Dis.* **2002**, *8* (9), 881–890.

(15) James, G. A.; Beaudette, L.; Costerton, J. W. Interspecies Bacterial Interactions in Biofilms. *J. Ind. Microbiol.* **1995**, *15* (4), 257–262. https://doi.org/10.1007/BF01569978.

(16) Gunn, J. S.; Bakaletz, L. O.; Wozniak, D. J. What's on the Outside Matters: The Role of the Extracellular Polymeric Substance of Gram-Negative Biofilms in Evading Host Immunity and as a Target for Therapeutic Intervention. *J. Biol. Chem.* **2016**, *291* (24), 12538–12546. https://doi.org/10.1074/jbc.R115.707547.

(17) Brauner, A.; Fridman, O.; Gefen, O.; Balaban, N. Q. Distinguishing between Resistance, Tolerance and Persistence to Antibiotic Treatment. *Nat. Rev. Microbiol.* **2016**, *14* (5), 320–330. https://doi.org/10.1038/nrmicro.2016.34.

(18) Koo, H.; Allan, R. N.; Howlin, R. P.; Stoodley, P.; Hall-Stoodley, L. Targeting Microbial Biofilms: Current and Prospective Therapeutic Strategies. *Nat. Rev. Microbiol.* **2017**, *15* (12), 740–755. https://doi.org/10.1038/nrmicro.2017.99.

(19) Hobby, G. L.; Karl, M.; Chaffee, E. Observations on the Mechanism of Action of Penicillin. *Exp. Biol. Med.* **1942**, *50* (2), 281–285.

(20) Bigger, J. W. Treatment of Staphylococcal Infections With Penicillin By Intermittent Sterilisation. *Lancet* **1944**, *244* (6320), 497–500. https://doi.org/10.1016/S0140-6736(00)74210-3.

(21) Lewis, K. Persister Cells, Dormancy and Infectious Disease. *Nat. Rev. Microbiol.* **2007**, *5* (1), 48–56. https://doi.org/10.1038/nrmicro1557.

(22) Flemming, H. C.; Wingender, J.; Szewzyk, U.; Steinberg, P.; Rice, S. A.; Kjelleberg, S. Biofilms: An Emergent Form of Bacterial Life. *Nature Reviews*

Microbiology. Nature Publishing Group 2016, pp 563–575. https://doi.org/10.1038/nrmicro.2016.94.

(23) Musk Jr., D.; Hergenrother, P. Chemical Countermeasures for the Control of Bacterial Biofilms: Effective Compounds and Promising Targets. *Curr. Med. Chem.* **2006**, *13* (18), 2163–2177. https://doi.org/10.2174/092986706777935212.

(24) Worthington, R. J.; Richards, J. J.; Melander, C. Small Molecule Control of Bacterial Biofilms. *Org. Biomol. Chem.* **2012**, *10* (37), 7457–7474. https://doi.org/10.1039/c2ob25835h.

(25) Wolcott, R. D.; Rhoads, D. D.; Bennett, M. E.; Wolcott, B. M.; Gogokhia, L.; Costerton, J. W.; Dowd, S. E. Chronic Wounds and the Medical Biofilm Paradigm. *J. Wound Care* **2010**, *19* (2), 45–53. https://doi.org/10.12968/jowc.2010.19.2.46966.

(26) American Thoracic Society; Infectious Diseases Society of America. Guidelines for the Management of Adults with Hospital-Acquired, Ventilator-Associated, and Healthcare-Associated Pneumonia. *Am. J. Respir. Crit. Care Med.* **2005**, *171* (4), 388–416. https://doi.org/10.1164/rccm.200405-644ST.

(27) Stickler, D. J. Bacterial Biofilms in Patients with Indwelling Urinary Catheters. *Nat. Clin. Pract. Urol.* **2008**, *5* (11), 598–608. https://doi.org/10.1038/ncpuro1231.

(28) Velkov, T.; Roberts, K. D.; Li, J. Rediscovering the Octapeptins. *Nat. Prod. Rep.* **2017**, *34* (3), 295–309. https://doi.org/10.1039/c6np00113k.

(29) Høiby, N.; Bjarnsholt, T.; Moser, C.; Bassi, G. L.; Coenye, T.; Donelli, G.; Hall-Stoodley, L.; Holá, V.; Imbert, C.; Kirketerp-Møller, K.; Lebeaux, D.; Oliver, A.; Ullmann, A. J.; Williams, C.; ESCMID Study Group for Biofilms (ESGB); Consulting External Expert Werner Zimmerli. ESCMID* Guideline for the Diagnosis and Treatment of Biofilm Infections 2014. *Clin. Microbiol. Infect.* **2015**, *21* (S1), S1–S25. https://doi.org/10.1016/j.cmi.2014.10.024.

(30) Raad, I.; Chaftari, A. M.; Zakhour, R.; Jordan, M.; Al Hamal, Z.; Jiang, Y.; Yousif, A.; Garoge, K.; Mulanovich, V.; Viola, G. M.; Kanj, S.; Pravinkumar, E.; Rosenblatt, J.; Hachem, R. Successful Salvage of Central Venous Catheters in Patients with Catheter-Related or Central Line-Associated Bloodstream Infections by Using a Catheter Lock Solution Consisting of Minocycline, EDTA,

and 25% Ethanol. *Antimicrob. Agents Chemother.* **2016**, *60* (6), 3426–3432. https://doi.org/10.1128/AAC.02565-15.

(31) Justo, J. A.; Bookstaver, P. B. Antibiotic Lock Therapy: Review of Technique and Logistical Challenges. *Infect. Drug Resist.* **2014**, *7*, 343–363. https://doi.org/10.2147/IDR.S51388.

(32) Blair, J. M. A.; Webber, M. A.; Baylay, A. J.; Ogbolu, D. O.; Piddock, L. J. V. Molecular Mechanisms of Antibiotic Resistance. *Nat. Publ. Gr.* **2014**, *13* (1), 42–51. https://doi.org/10.1038/nrmicro3380.

(33) Banerjee, G.; Ray, A. K. Quorum-Sensing Network-Associated Gene Regulation in Gram-Positive Bacteria. *Acta Microbiol. Immunol. Hung.* **2017**, *64* (4), 439–453. https://doi.org/10.1556/030.64.2017.040.

(34) Nealson, K. H.; Hastings, J. W. Bacterial Bioluminescence: Its Control and Ecological Significance. *Microbiol. Rev.* **1979**, *43* (4), 496–518. https://doi.org/10.1128/mmbr.43.4.496-518.1979.

(35) Fuqua, C.; Parsek, M. R.; Greenberg, E. P. Regulation of Gene Expression by Cell-to-Cell Communication: Acyl-Homoserine Lactone Quorum Sensing. *Annu. Rev. Genet.* **2001**, *35* (1), 439–468. https://doi.org/10.1146/annurev.genet.35.102401.090913.

(36) Sankar Ganesh, P.; Ravishankar Rai, V. Attenuation of Quorum-Sensing-Dependent Virulence Factors and Biofilm Formation by Medicinal Plants against Antibiotic Resistant Pseudomonas Aeruginosa. *J. Tradit. Complement. Med.* **2018**, *8* (1), 170–177. https://doi.org/10.1016/j.jtcme.2017.05.008.

(37) Le Berre, R.; Nguyen, S.; Nowak, E.; Kipnis, E.; Pierre, M.; Quenee, L.; Ader, F.; Lancel, S.; Courcol, R.; Guery, B. P.; Faure, K. Relative Contribution of Three Main Virulence Factors in Pseudomonas Aeruginosa Pneumonia. *Crit. Care Med.* **2011**, *39* (9), 2113–2120. https://doi.org/10.1097/CCM.0b013e31821e899f.

(38) Gallardo-García, M. M.; Sánchez-Espín, G.; Ivanova-Georgieva, R.; Ruíz-Morales, J.; Rodríguez-Bailón, I.; Viñuela González, V.; García-López, M. V. Relationship between Pathogenic, Clinical, and Virulence Factors of Staphylococcus Aureus in Infective Endocarditis versus Uncomplicated Bacteremia: A Case–Control Study. *Eur. J. Clin. Microbiol. Infect. Dis.* **2016**, *35* (5), 821–828. https://doi.org/10.1007/s10096-016-2603-2.

(39) Mattmann, M. E.; Blackwell, H. E. Small Molecules That Modulate Quorum

Sensing and Control Virulence in Pseudomonas Aeruginosa. *J. Org. Chem.* **2010**, *75* (20), 6737–6746. https://doi.org/10.1021/jo101237e.

(40) Abraham, W.-R. Controlling Biofilms of Gram-Positive Pathogenic Bacteria. *Curr. Med. Chem.* **2006**, *13* (13), 1509–1524. https://doi.org/10.2174/092986706777442039.

(41) Jiang, Q.; Chen, J.; Yang, C.; Yin, Y.; Yao, K.; Song, D. Quorum Sensing: A Prospective Therapeutic Target for Bacterial Diseases. *Biomed Res. Int.* **2019**, *2019*. https://doi.org/10.1155/2019/2015978.

(42) Parsek, M. R.; Greenberg, E. P. Acyl-Homoserine Lactone Quorum Sensing in Gram-Negative Bacteria: A Signaling Mechanism Involved in Associations with Higher Organisms. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (16), 8789–8793. https://doi.org/10.1073/pnas.97.16.8789.

(43) Takano, E.; Chakraburtty, R.; Nihira, T.; Yamada, Y.; Bibb, M. J. A Complex Role for the γ-Butyrolactone SCB1 in Regulating Antibiotic Production in Streptomyces Coelicolor A3(2). *Mol. Microbiol.* **2001**, *41* (5), 1015–1028. https://doi.org/10.1046/j.1365-2958.2001.02562.x.

(44) Lyon, G. J.; Muir, T. W. Chemical Signaling among Bacteriaand Its Inhibition. *Chem. Biol.* **2003**, *10* (11), 1007–1021. https://doi.org/10.1016/j.

(45) Waters, C. M.; Bassler, B. L. QUORUM SENSING: Cell-to-Cell Communication in Bacteria. *Annu. Rev. Cell Dev. Biol.* **2005**, *21* (1), 319–346. https://doi.org/10.1146/annurev.cellbio.21.012704.131001.

(46) Mangwani, N.; Kumari, S.; Das, S. Bacterial Biofilms and Quorum Sensing: Fidelity in Bioremediation Technology. *Biotechnol. Genet. Eng. Rev.* **2017**, *32* (1–2), 43–73. https://doi.org/10.1080/02648725.2016.1196554.

(47) Davies, D. G.; Marques, C. N. H. A Fatty Acid Messenger Is Responsible for Inducing Dispersion in Microbial Biofilms. *J. Bacteriol.* **2009**, *191* (5), 1393–1403. https://doi.org/10.1128/JB.01214-08.

(48) Dow, J. M.; Crossman, L.; Findlay, K.; He, Y.; Feng, J.; Tang, J. Biofilm Dispersal in Xanthomonas Campestris Is Controlled by Cell– Cell Signaling and Is Required for Full Virulence to Plants. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (19).

(49) Kolodkin-gal, I.; Romero, D.; Cao, S.; Clardy, J.; Kolter, R.; Losick, R. D-Amino Acids Trigger Biofilm Disassembly. *Science (80-.).* **2010**, *328* (April), 627–630.

https://doi.org/10.7551/mitpress/8876.003.0036.

(50) Götz, F. Staphylococcus and Biofilms. *Mol. Microbiol.* **2002**, *43* (6), 1367–1378. https://doi.org/10.1046/j.1365-2958.2002.02827.x.

(51) Fux, C. A.; Stoodley, P.; Hall-Stoodley, L.; Costerton, J. W. Bacterial Biofilms: A Diagnostic and Therapeutic Challenge. *Expert Rev. Anti. Infect. Ther.* **2003**, *1* (4), 667–683. https://doi.org/10.1586/14787210.1.4.667.

(52) Yarwood, J. M.; Bartels, D. J.; Volper, E. M.; Greenberg, E. P. Quorum Sensing in Staphylococcus Aureus Biofilms. *J. Bacteriol.* **2004**, *186* (6), 1838–1850. https://doi.org/10.1128/JB.186.6.1838-1850.2004.

(53) Fazeli, H.; Akbari, R.; Moghim, S.; Mohammad, A.; Ghoddous, A. Pseudomonas Aeruginosa Infections in Patients, Hospital Means, and Personnel's Specimens. *J. Res. Med. Sci.* **2012**, *17* (4), 332–337.

(54) Sadikot, R. T.; Blackwell, T. S.; Christman, J. W.; Prince, A. S. Pathogen-Host Interactions in Pseudomonas Aeruginosa Pneumonia. *Am. J. Respir. Crit. Care Med.* **2005**, *171* (11), 1209–1223. https://doi.org/10.1164/rccm.200408-1044SO.

(55) Skariyachan, S.; Sridhar, V. S.; Packirisamy, S.; Kumargowda, S. T.; Challapilli, S. B. Recent Perspectives on the Molecular Basis of Biofilm Formation by Pseudomonas Aeruginosa and Approaches for Treatment and Biofilm Dispersal. *Folia Microbiol. (Praha).* **2018**, *63* (4), 413–432. https://doi.org/10.1007/s12223-018-0585-4.

(56) Lee, J.; Zhang, L. The Hierarchy Quorum Sensing Network in Pseudomonas Aeruginosa. *Protein Cell* **2014**, *6* (1), 26–41. https://doi.org/10.1007/s13238-014-0100-x.

(57) Kumar, Mr. Chromobacterium Violaceum : A Rare Bacterium Isolated from a Wound over the Scalp. *Int. J. Appl. Basic Med. Res.* **2012**, *2* (1), 70. https://doi.org/10.4103/2229-516x.96814.

(58) Yang, C. H.; Li, Y. H. Chromobacterium Violaceum Infection: A Clinical Review of an Important but Neglected Infection. *J. Chinese Med. Assoc.* **2011**, *74* (10), 435–441. https://doi.org/10.1016/j.jcma.2011.08.013.

(59) Batista, J. H.; Neto, J. F. d. S. Chromobacterium Violaceum Pathogenicity: Updates and Insights from Genome Sequencing of Novel Chromobacterium Species. *Front. Microbiol.* **2017**, *8* (NOV), 1–7. https://doi.org/10.3389/fmicb.2017.02213.

(60) McClean, K. H.; Winson, M. K.; Fish, L.; Taylor, A.; Chhabra, S. R.; Camara, M.; Daykin, M.; Lamb, J. H.; Swift, S.; Bycroft, B. W.; Stewart, G. S. A. B.; Williams, P. Quorum Sensing and Chromobacterium Violaceum: Exploitation of Violacein Production and Inhibition for the Detection of N-Acylhomoserine Lactones. *Microbiology* **1997**, *143* (12), 3703–3711. https://doi.org/10.1099/00221287-143-12-3703.

(61) Kothari, V.; Sharma, S.; Padia, D. Recent Research Advances on Chromobacterium Violaceum. *Asian Pac. J. Trop. Med.* **2017**, *10* (8), 744–752. https://doi.org/10.1016/j.apjtm.2017.07.022.

(62) Chernin, L. S.; Winson, M. K.; Thompson, J. M.; Haran, S.; Bycroft, B. W.; Chet, I.; Williams, P.; Stewart, G. S. A. B. Chitinolytic Activity in Chromobacterium Violaceum: Substrate Analysis and Regulation by Quorum Sensing. *J. Bacteriol.* **1998**, *180* (17), 4435–4441. https://doi.org/10.1128/jb.180.17.4435-4441.1998.

(63) Morohoshi, T.; Kato, M.; Fukamachi, K.; Kato, N.; Ikeda, T. N-Acylhomoserine Lactone Regulates Violacein Production in Chromobacterium Violaceum Type Strain ATCC 12472. *FEMS Microbiol. Lett.* **2008**, *279* (1), 124–130. https://doi.org/10.1111/j.1574-6968.2007.01016.x.

(64) Hoshino, T. Violacein and Related Tryptophan Metabolites Produced by Chromobacterium Violaceum: Biosynthetic Mechanism and Pathway for Construction of Violacein Core. *Appl. Microbiol. Biotechnol.* **2011**, *91* (6), 1463–1475. https://doi.org/10.1007/s00253-011-3468-z.

(65) Stauff, D. L.; Bassler, B. L. Quorum Sensing in Chromobacterium Violaceum: DNA Recognition and Gene Regulation by the CviR Receptor. *J. Bacteriol.* **2011**, *193* (15), 3871–3878. https://doi.org/10.1128/JB.05125-11.

(66) Hanzelka, B. L.; Greenberg, E. P. Evidence That the N-Terminal Region of the Vibrio Fischeri LuxR Protein Constitutes an Autoinducer-Binding Domain. *J. Bacteriol.* **1995**, *177* (3), 815–817. https://doi.org/10.1128/jb.177.3.815-817.1995.

(67) Chen, G.; Swem, L. R.; Swem, D. L.; Stauff, D. L.; O'Loughlin, C. T.; Jeffrey, P. D.; Bassler, B. L.; Hughson, F. M. A Strategy for Antagonizing Quorum Sensing. *Mol. Cell* **2011**, *42* (2), 199–209. https://doi.org/10.1016/j.molcel.2011.04.003.

(68) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.;

Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. The Protein Data Bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2002**, *58* (6 I), 899–907. https://doi.org/10.1107/S0907444902003451.

(69) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (D1), 1100–1107. https://doi.org/10.1093/nar/gkr777.

(70) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44* (D1), D1202–D1213. https://doi.org/10.1093/nar/gkv951.

(71) Galperin, M. Y. The Molecular Biology Database Collection: 2005 Update. *Nucleic Acids Res.* **2005**, *33* (DATABASE ISS.). https://doi.org/10.1093/nar/gki139.

(72) Rigden, D. J.; Fernández, X. M. The 27th Annual Nucleic Acids Research Database Issue and Molecular Biology Database Collection. *Nucleic Acids Res.* **2020**, *48* (D1), D1–D8. https://doi.org/10.1093/nar/gkz1161.

(73) Wynendaele, E.; Bronselaer, A.; Nielandt, J.; D'Hondt, M.; Stalmans, S.; Bracke, N.; Verbeke, F.; Van De Wiele, C.; De Tré, G.; De Spiegeleer, B. Quorumpeps Database: Chemical Space, Microbial Origin and Functionality of Quorum Sensing Peptides. *Nucleic Acids Res.* **2013**, *41* (D1), 655–659. https://doi.org/10.1093/nar/gks1137.

(74) Lourenço, A.; Ferreira, A.; Veiga, N.; Machado, I.; Pereira, M. O.; Azevedo, N. F. Biofomics: A Web Platform for the Systematic and Standardized Collection of High-Throughput Biofilm Data. *PLoS One* **2012**, *7* (6). https://doi.org/10.1371/journal.pone.0039960.

(75) Di Luca, M.; Maccari, G.; Maisetta, G.; Batoni, G. BaAMPs: The Database of Biofilm-Active Antimicrobial Peptides. *Biofouling* **2015**, *31* (2), 193–199. https://doi.org/10.1080/08927014.2015.1021340.

(76) Rajput, A.; Gupta, A. K.; Kumar, M. Prediction and Analysis of Quorum Sensing Peptides Based on Sequence Features. *PLoS One* **2015**, *10* (3), 1–16. https://doi.org/10.1371/journal.pone.0120066.

(77) Sharma, A.; Gupta, P.; Kumar, R.; Bhardwaj, A. DPABBs: A Novel in Silico

Approach for Predicting and Designing Anti-Biofilm Peptides. *Sci. Rep.* **2016**, *6* (July 2015), 1–13. https://doi.org/10.1038/srep21839.

(78) Rajput, A.; Thakur, A.; Sharma, S.; Kumar, M. ABiofilm: A Resource of Anti-Biofilm Agents and Their Potential Implications in Targeting Antibiotic Drug Resistance. *Nucleic Acids Res.* **2018**, *46* (D1), D894–D900. https://doi.org/10.1093/nar/gkx1157.

(79) Magalhães, R. P.; Vieira, T. F.; Fernandes, H. S.; Melo, A.; Simões, M.; Sousa, S. F. The Biofilms Structural Database. *Trends Biotechnol.* **2020**, *xx* (xx), 1–4. https://doi.org/10.1016/j.tibtech.2020.04.002.

(80) Kaur, A.; Capalash, N.; Sharma, P. Quorum Sensing in Thermophiles: Prevalence of Autoinducer-2 System. *BMC Microbiol.* **2018**, *18* (1), 1–16. https://doi.org/10.1186/s12866-018-1204-x.

(81) Rajput, A.; Kumar, M. In Silico Analyses of Conservational, Functional and Phylogenetic Distribution of the LuxI and LuxR Homologs in Gram-Positive Bacteria. *Sci. Rep.* **2017**, *7* (1), 1–13. https://doi.org/10.1038/s41598-017-07241-5.

(82) Cattò, C.; Cappitelli, F. Testing Anti-Biofilm Polymeric Surfaces: Where to Start? *Int. J. Mol. Sci.* **2019**, *20* (15). https://doi.org/10.3390/ijms20153794.

(83) Boudarel, H.; Mathias, J. D.; Blaysat, B.; Grédiac, M. Towards Standardized Mechanical Characterization of Microbial Biofilms: Analysis and Critical Review. *npj Biofilms Microbiomes* **2018**, *4* (1). https://doi.org/10.1038/s41522-018-0062-5.

(84) Azeredo, J.; Azevedo, N. F.; Briandet, R.; Cerca, N.; Coenye, T.; Costa, A. R.; Desvaux, M.; Di Bonaventura, G.; Hébraud, M.; Jaglic, Z.; Kačániová, M.; Knøchel, S.; Lourenço, A.; Mergulhão, F.; Meyer, R. L.; Nychas, G.; Simões, M.; Tresse, O.; Sternberg, C. Critical Review on Biofilm Methods. *Crit. Rev. Microbiol.* **2017**, *43* (3), 313–351. https://doi.org/10.1080/1040841X.2016.1208146.

(85) Hancock, R. E. W.; Sahl, H. G. Antimicrobial and Host-Defense Peptides as New Anti-Infective Therapeutic Strategies. *Nat. Biotechnol.* **2006**, *24* (12), 1551–1557. https://doi.org/10.1038/nbt1267.

(86) Gupta, S.; Sharma, A. K.; Jaiswal, S. K.; Sharma, V. K. Prediction of Biofilm Inhibiting Peptides: An In Silico Approach. *Front. Microbiol.* **2016**, *7* (JUN), 1–11.

https://doi.org/10.3389/fmicb.2016.00949.

(87) Vergis, J.; Malik, S. S.; Pathak, R.; Kumar, M.; Ramanjaneya, S.; Kurkure, N. V.; Barbuddhe, S. B.; Rawool, D. B. Antimicrobial Efficacy of Indolicidin Against Multi-Drug Resistant Enteroaggregative Escherichia Coli in a Galleria Mellonella Model. *Front. Microbiol.* **2019**, *10* (November). https://doi.org/10.3389/fmicb.2019.02723.

(88) Moussa, D. G.; Fok, A.; Aparicio, C. Hydrophobic and Antimicrobial Dentin: A Peptide-Based 2-Tier Protective System for Dental Resin Composite Restorations. *Acta Biomater.* **2019**, *88*, 251–265. https://doi.org/10.1016/j.actbio.2019.02.007.

(89) Pang, X.; Liu, C.; Lv, P.; Zhang, S.; Liu, L.; Lu, J.; Lv, J. Identification of Quorum Sensing Signal Molecule of Lactobacillus Delbrueckii Subsp . Bulgaricus Identification of Quorum Sensing Signal Molecule of Lactobacillus Delbrueckii Subsp . Bulgaricus. *J. Agric. Food Chem* **2016**, *64* (39), 9421–9427. https://doi.org/10.1021/acs.jafc.6b04016.

(90) Chegini, P. P.; Nikokar, I.; Tabarzad, M.; Faezi, S.; Mahboubi, A. Effect of Amino Acid Substitutions on Biological Activity of Antimicrobial Peptide: Design, Recombinant Production, and Biological Activity. *Iran. J. Pharm. Res.* **2019**, *18* (Special Issue), 157–168. https://doi.org/10.22037/ijpr.2019.112397.13734.

(91) Marimuthu, S. K.; Nagarajan, K.; Perumal, S. K.; Palanisamy, S.; Subbiah, L. Insilico Alpha-Helical Structural Recognition of Temporin Antimicrobial Peptides and Its Interactions with Middle East Respiratory Syndrome-Coronavirus. *Int. J. Pept. Res. Ther.* **2020**, *26* (3), 1473–1483. https://doi.org/10.1007/s10989-019-09951-y.

(92) Leoni, G.; De Poli, A.; Mardirossian, M.; Gambato, S.; Florian, F.; Venier, P.; Wilson, D. N.; Tossi, A.; Pallavicini, A.; Gerdol, M. Myticalins: A Novel Multigenic Family of Linear, Cationic Antimicrobial Peptides from Marine Mussels (Mytilus Spp.). *Mar. Drugs* **2017**, *15* (8), 1–23. https://doi.org/10.3390/md15080261.

(93) Tiwari, V.; Meena, K.; Tiwari, M. Differential Anti-Microbial Secondary Metabolites in Different ESKAPE Pathogens Explain Their Adaptation in the Hospital Setup. *Infect. Genet. Evol.* **2018**, *66* (June), 57–65. https://doi.org/10.1016/j.meegid.2018.09.010.

(94) Almeida, F. A. de; Vargas, E. L. G.; Carneiro, D. G.; Pinto, U. M.; Vanetti, M. C.

D. Virtual Screening of Plant Compounds and Nonsteroidal Anti-Inflammatory Drugs for Inhibition of Quorum Sensing and Biofilm Formation in Salmonella. *Microb. Pathog.* **2018**, *121* (April), 369–388. https://doi.org/10.1016/j.micpath.2018.05.014.

(95) Sun, Y. Z.; Zhang, D. H.; Cai, S. Bin; Ming, Z.; Li, J. Q.; Chen, X. MDAD: A Special Resource for Microbe-Drug Associations. *Front. Cell. Infect. Microbiol.* **2018**, *8* (December), 424. https://doi.org/10.3389/fcimb.2018.00424.

(96) Vieira, T. F.; Sousa, S. F. Comparing AutoDock and Vina in Ligand/Decoy Discrimination for Virtual Screening. *Appl. Sci.* **2019**, *9* (21). https://doi.org/10.3390/app9214538.

(97) F. Sousa, S.; M.F.S.A. Cerqueira, N.; A. Fernandes, P.; Joao Ramos, M. Virtual Screening in Drug Design and Development. *Comb. Chem. High Throughput Screen.* **2010**, *13* (5), 442–453. https://doi.org/10.2174/138620710791293001.

(98) Gimeno, A.; Ojeda-Montes, M. J.; Tomás-Hernández, S.; Cereto-Massagué, A.; Beltrán-Debón, R.; Mulero, M.; Pujadas, G.; Garcia-Vallvé, S. The Light and Dark Sides of Virtual Screening: What Is There to Know? *Int. J. Mol. Sci.* **2019**, *20* (6). https://doi.org/10.3390/ijms20061375.

(99) Neves, R. P. P.; Sousa, S. F.; Ribeiro, A. J. M.; Coimbra, J. T. S.; Martins, S. A.; Moorthy, N. S. H. N.; Fernandes, P. A.; Ramos, M. J. Protein-Ligand Docking in the New Millennium – A Retrospective of 10 Years in the Field Protein-Ligand Docking in the New Millennium – A Retrospective of 10 Years in the Field. *Curr. Med. Chem.* **2013**, *20* (18), 2296–2314. https://doi.org/10.2174/0929867311320180002.

(100) Grinter, S. Z.; Zou, X. Challenges, Applications, and Recent Advances of Protein-Ligand Docking in Structure-Based Drug Design. *Molecules* **2014**, *19* (7), 10150–10176. https://doi.org/10.3390/molecules190710150.

(101) Waszkowycz, B.; Clark, D. E.; Gancia, E. Outstanding Challenges in Protein-Ligand Docking and Structure-Based Virtual Screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (2), 229–259. https://doi.org/10.1002/wcms.18.

(102) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. QSAR without Borders. *Chem. Soc.*

Rev. **2020**, *49* (11), 3525–3564. https://doi.org/10.1039/d0cs00098a.

(103) Toropov, A. A.; Toropova, A. P. QSPR/QSAR: State-of-Art,Weirdness, the Future. *Molecules* **2020**, *25* (6). https://doi.org/10.3390/molecules25061292.

(104) Ganesan, A.; Coote, M. L.; Barakat, K. Molecular Dynamics-Driven Drug Discovery: Leaping Forward with Confidence. *Drug Discov. Today* **2017**, *22* (2), 249–269. https://doi.org/10.1016/j.drudis.2016.11.001.

(105) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99* (6), 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011.

(106) Wen, B.; Liu, J.-H.; Zhang, Y.; Zhang, H.-R.; Gao, J.-Z.; Chen, Z.-Z. Community Structure and Functional Diversity of the Plastisphere in Aquaculture Waters: Does Plastic Color Matter? *Sci. Total Environ.* **2020**, 140082. https://doi.org/10.1016/j.scitotenv.2020.140082.

(107) Flemming, H. C. Biofouling and Me: My Stockholm Syndrome with Biofilms. *Water Res.* **2020**, *173*, 2020. https://doi.org/10.1016/j.watres.2020.115576.

(108) Cho, K. H.; Tryon, R. G.; Kim, J. H. Screening for Diguanylate Cyclase (DGC) Inhibitors Mitigating Bacterial Biofilm Formation. *Front. Chem.* **2020**, *8*, 2020. https://doi.org/10.3389/fchem.2020.00264.

(109) Tobal, I. E.; Roncero, A. M.; Moro, R. F.; Díez, D.; Marcos, I. S. Antibacterial Natural Halimanes: Potential Source of Novel Antibiofilm Agents. *Molecules* **2020**, *25* (7), 2020. https://doi.org/10.3390/molecules25071707.

(110) Pedroza-Dávila, U.; Uribe-Alvarez, C.; Morales-García, L.; Espinoza-Simón, E.; Méndez-Romero, O.; Muhlia-Almazán, A.; Chiquete-Félix, N.; Uribe-Carvajal, S. Metabolism, ATP Production and Biofilm Generation by Staphylococcus Epidermidis in Either Respiratory or Fermentative Conditions. *AMB Express* **2020**, *10* (1), 2020. https://doi.org/10.1186/s13568-020-00966-z.

(111) Machineni, L. Effects of Biotic and Abiotic Factors on Biofilm Growth Dynamics and Their Heterogeneous Response to Antibiotic Challenge. *J. Biosci.* **2020**, *45* (1), 2020. https://doi.org/10.1007/s12038-020-9990-3.

(112) Thöming, J. G.; Tomasch, J.; Preusse, M.; Koska, M.; Grahl, N.; Pohl, S.; Willger, S. D.; Kaever, V.; Müsken, M.; Häussler, S. Parallel Evolutionary Paths to Produce More than One Pseudomonas Aeruginosa Biofilm Phenotype. *npj Biofilms Microbiomes* **2020**, *6* (1), 2020. https://doi.org/10.1038/s41522-019-0113-6.

(113) Cui, Y.; Schmid, B. V.; Cao, H.; Dai, X.; Du, Z.; Ryan Easterday, W.; Fang, H.; Guo, C.; Huang, S.; Liu, W.; Qi, Z.; Song, Y.; Tian, H.; Wang, M.; Wu, Y.; Xu, B.; Yang, C.; Yang, J.; Yang, X.; Zhang, Q.; Jakobsen, K. S.; Zhang, Y.; Stenseth, N. C.; Yang, R. Evolutionary Selection of Biofilm-Mediated Extended Phenotypes in Yersinia Pestis in Response to a Fluctuating Environment. *Nat. Commun.* **2020**, *11* (1), 2020. https://doi.org/10.1038/s41467-019-14099-w.

(114) Nelson, K. S.; Baltar, F.; Lamare, M. D.; Morales, S. E. Ocean Acidification Affects Microbial Community and Invertebrate Settlement on Biofilms. *Sci. Rep.* **2020**, *10* (1), 2020. https://doi.org/10.1038/s41598-020-60023-4.

(115) Peng, L. H.; Liang, X.; Xu, J. K.; Dobretsov, S.; Yang, J. L. Monospecific Biofilms of Pseudoalteromonas Promote Larval Settlement and Metamorphosis of Mytilus Coruscus. *Sci. Rep.* **2020**, *10* (1), 2020. https://doi.org/10.1038/s41598-020-59506-1.

(116) Grudlewska-Buda, K.; Skowron, K.; Gospodarek-Komkowska, E. Comparison of the Intensity of Biofilm Formation by Listeria Monocytogenes Using Classical Culture-Based Method and Digital Droplet PCR. *AMB Express* **2020**, *10* (1), 2020. https://doi.org/10.1186/s13568-020-01007-5.

(117) Bamford, N. C.; Le Mauff, F.; Van Loon, J. C.; Ostapska, H.; Snarr, B. D.; Zhang, Y.; Kitova, E. N.; Klassen, J. S.; Codée, J. D. C.; Sheppard, D. C.; Howell, P. L. Structural and Biochemical Characterization of the Exopolysaccharide Deacetylase Agd3 Required for Aspergillus Fumigatus Biofilm Formation. *Nat. Commun.* **2020**, *11* (1), 2020. https://doi.org/10.1038/s41467-020-16144-5.

(118) Zhao, F.; Yang, H.; Bi, D.; Khaledi, A.; Qiao, M. A Systematic Review and Meta-Analysis of Antibiotic Resistance Patterns, and the Correlation between Biofilm Formation with Virulence Factors in Uropathogenic E. Coli Isolated from Urinary Tract Infections. *Microb. Pathog.* **2020**, *144*, 2020. https://doi.org/10.1016/j.micpath.2020.104196.

(119) Gericke, B.; Schecker, N.; Amiri, M.; Naim, H. Y. Structure-Function Analysis of Human Sucrase-Isomaltase Identifies Key Residues Required for Catalytic Activity. **2017**, No. 4. https://doi.org/10.1074/jbc.M117.791939.

(120) Vyas, N.; Wang, Q. X.; Manmi, K. A.; Sammons, R. L.; Kuehne, S. A.; Walmsley, A. D. How Does Ultrasonic Cavitation Remove Dental Bacterial Biofilm? *Ultrason. Sonochem.* **2020**, *67*, 105112. https://doi.org/10.1016/j.ultsonch.2020.105112.

(121) Jeon, D. M.; An, J. S.; Lim, B. S.; Ahn, S. J. Orthodontic Bonding Procedures Significantly Influence Biofilm Composition. *Prog. Orthod.* **2020**, *21* (1), 2020. https://doi.org/10.1186/s40510-020-00314-8.

(122) Lehninger, A. L.; Nelson, D. L.; Cox, M. M. *Lehninger Principles of Biochemistry. Seventh Edition*, 7th ed.; Macmillan Higher Education, 2017.

(123) Du, X.; Li, Y.; Xia, Y. L.; Ai, S. M.; Liang, J.; Sang, P.; Ji, X. L.; Liu, S. Q. Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods. *Int. J. Mol. Sci.* **2016**, *17* (2), 1–34. https://doi.org/10.3390/ijms17020144.

(124) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve RD Productivity: The Pharmaceutical Industry's Grand Challenge. *Nat. Rev. Drug Discov.* **2010**, *9* (3), 203–214. https://doi.org/10.1038/nrd3078.

(125) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66* (1), 334–395. https://doi.org/10.1124/pr.112.007336.

(126) Macalino, S. J. Y.; Gosu, V.; Hong, S.; Choi, S. Role of Computer-Aided Drug Design in Modern Drug Discovery. *Arch. Pharm. Res.* **2015**, *38* (9), 1686–1701. https://doi.org/10.1007/s12272-015-0640-5.

(127) Bylinsky, G. A New Industrial Revolution Is on the Way. *Fortune*. 1981, pp 106–114.

(128) Cerqueira, N. M. F. S. A.; Gesto, D.; Oliveira, E. F.; Santos-Martins, D.; Brás, N. F.; Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Receptor-Based Virtual Screening Protocol for Drug Discovery. *Arch. Biochem. Biophys.* **2015**, *582*, 56–67. https://doi.org/10.1016/j.abb.2015.05.011.

(129) Meng, X.; Zhang, H.; Mezei, M.; Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr. Comput. Aided-Drug Des.* **2012**, *7* (2), 146–157. https://doi.org/10.2174/157340911795677602.

(130) Brás, N. F.; Cerqueira, N. M. F. S. A.; Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein Ligand Docking in Drug Discovery. *Protein Model.* **2014**, *9783319099*, 249–286. https://doi.org/10.1007/978-3-319-09976-7_11.

(131) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30* (16), 2785–2791.

https://doi.org/10.1002/jcc.21256.AutoDock4.

(132) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19* (14), 1639–1662.

(133) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A Semiempirical Free Energy Force Field with Charge-Based Desolvation. *J. Comput. Chem.* **2007**, No. 28, 1145–1152. https://doi.org/10.1002/jcc.

(134) Trott, O.; Olson, A. J. Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31* (2), 455–461. https://doi.org/10.1002/jcc.

(135) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267* (3), 727–748.

(136) Mooij, W. T. M.; Verdonk, M. L. General and Targeted Statistical Potentials for Protein-Ligand Interactions. *Proteins Struct. Funct. Genet.* **2005**, *61* (2), 272–287. https://doi.org/10.1002/prot.20588.

(137) Korb, O.; Stützle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49* (1), 84–96. https://doi.org/10.1021/ci800298z.

(138) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein–Ligand Docking Using GOLD. *Proteins* **2003**, *52* (January), 609–623. https://doi.org/10.1002/prot.10465.

(139) 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.* **1997**, *11* (5), 425–445. https://doi.org/10.1023/A:1007996124545.

(140) Zhang, N.; Zhao, H. Enriching Screening Libraries with Bioactive Fragment Space. *Bioorg. Med. Chem. Lett.* **2016**, *26* (15), 3594–3597. https://doi.org/10.1016/j.bmcl.2016.06.013.

(141) Gaulton, A.; Hersey, A.; Nowotka, M. L.; Patricia Bento, A.; Chambers, J.;

Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magarinos, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45* (D1), D945–D954. https://doi.org/10.1093/nar/gkw1074.

(142) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594. https://doi.org/10.1021/jm300687e.

(143) Empereur-Mot, C.; Zagury, J. F.; Montes, M. Screening Explorer-An Interactive Tool for the Analysis of Screening Results. *J. Chem. Inf. Model.* **2016**, *56* (12), 2281–2286. https://doi.org/10.1021/acs.jcim.6b00283.

(144) Truchon, J. F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the “Early Recognition” Problem. *J. Chem. Inf. Model.* **2007**, *47* (2), 488–508. https://doi.org/10.1021/ci600426e.

(145) Leach, A. R. *Molecular Modelling: Principles and Applications*, Second.; Prentice Hall, 2001.

(146) Adcock, S. A.; Mccammon, J. A. Molecular Dynamics : Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* **2006**, *106* (5), 1589–1615.

(147) Ponder, J. W.; Case, D. A. Force Fields for Protein Simulations. *Adv. Protein Chem.* **2003**, *66*, 27–85. https://doi.org/10.1016/S0065-3233(03)66002-X.

(148) Kräutler, V.; Van Gunsteren, W. F.; Hünenberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. *J. Comput. Chem.* **2001**, *22* (5), 501–508. https://doi.org/10.1002/1096-987X(20010415)22:5<501::AID-JCC1021>3.0.CO;2-V.

(149) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Fergunson, David, M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179–5197.

(150) Case, D. A.; Cheatham, T. E.; Darden, T. O. M.; Gohlke, H.; Luo, R. A. Y.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688. https://doi.org/10.1002/jcc.20290.

(151) Do Vale Hipolito Cavalheiro, J. P.; Pires, N. M. M.; Dong, T. MM-PBSA: Challenges and Opportunities. *Proc. - 2017 10th Int. Congr. Image Signal Process. Biomed. Eng. Informatics, CISP-BMEI 2017* **2018**, *2018-Janua* (1), 1–6. https://doi.org/10.1109/CISP-BMEI.2017.8302303.

(152) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opin. Drug Discov.* **2015**, *10* (5), 449–461. https://doi.org/10.1517/17460441.2015.1032936.

(153) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA.Py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8* (9), 3314–3321. https://doi.org/10.1021/ct300418h.

(154) Durham, E.; Dorr, B.; Woetzel, N.; Staritzbichler, R.; Meiler, J. Solvent Accessible Surface Area Approximations for Rapid and Accurate Protein Structure Prediction. *J. Mol. Model.* **2009**, *15* (9), 1093–1108. https://doi.org/10.1007/s00894-009-0454-9.

(155) DeLano, W. L. Pymol: An Open-Source Molecular Graphics Tool. *CCP4 Newsl. Protein Crystallogr.* **2002**, *40*, 82–92.

(156) Lazakidou, A. *Biocomputation and Biomedical Informatics: Case Studies and Applications*, First Edit.; Lazakidou, A., Ed.; Medical Information Science Reference, 2009.

(157) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38.

(158) Westbrook, J. D.; Fitzgerald, P. M. D. The PDB Format, MmCIF Formats, and Other Data Formats. In *Structural Bioinformatics, Volume 44*; Bourne, P. E., Weissig, H., Eds.; John Wiley & Sons, Inc., 2003; pp 161–179.

(159) AutoDock FAQ: What is the format of a PDBQT file? http://autodock.scripps.edu/faqs-help/faq/what-is-the-format-of-a-pdbqt-file (accessed Apr 20, 2020).

(160) Boyle, N. M. O.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3* (33).

(161) Bell, E. W.; Zhang, Y. DockRMSD: An Open-Source Tool for Atom Mapping and RMSD Calculation of Symmetric Molecules through Graph Isomorphism. *J.*

Cheminform. **2019**, *11* (1), 1–9. https://doi.org/10.1186/s13321-019-0362-7.

(162) Quecan, B. X. V.; Santos, J. T. C.; Rivera, M. L. C.; Hassimotto, N. M. A.; Almeida, F. A.; Pinto, U. M. Effect of Quercetin Rich Onion Extracts on Bacterial Quorum Sensing. *Front. Microbiol.* **2019**, *10* (APR), 1–16. https://doi.org/10.3389/fmicb.2019.00867.

(163) Sun, S.; Dai, X.; Sun, J.; Bu, X.; Weng, C.; Li, H.; Zhu, H. A Diketopiperazine Factor from Rheinheimera Aquimaris QSI02 Exhibits Anti-Quorum Sensing Activity. *Sci. Rep.* **2016**, *6* (July), 1–10. https://doi.org/10.1038/srep39637.

(164) Rivera, M. L. C.; Hassimotto, N. M. A.; Bueris, V.; Sircili, M. P.; de Almeida, F. A.; Pinto, U. M. Effect of Capsicum Frutescens Extract, Capsaicin, and Luteolin on Quorum Sensing Regulated Phenotypes. *J. Food Sci.* **2019**, *84* (6), 1477–1486. https://doi.org/10.1111/1750-3841.14648.

(165) Singh, S.; Bhatia, S. In Silico Identification of Albendazole as a Quorum Sensing Inhibitor and Its in Vitro Verification Using CviR and LasB Receptors Based Assay Systems. *BioImpacts* **2018**, *8* (3), 201–209. https://doi.org/10.15171/bi.2018.23.

(166) Juan, V.; Ramírez-ch, E.; Gutierrez-villagomez, J. M.; García-gonz, J. P.; Molina-torres, J. Bioautography and GC-MS Based Identification of Piperine and Trichostachine as the Active Quorum Quenching Compounds in Black Pepper. **2020**, *6* (November 2019). https://doi.org/10.1016/j.heliyon.2019.e03137.

(167) Kimyon, Ö.; Ulutürk, Z. I.; Nizalapur, S.; Lee, M.; Kutty, S. K.; Beckmann, S.; Kumar, N.; Manefield, M. N-Acetylglucosamine Inhibits LuxR, LasR and CviR Based Quorum Sensing Regulated Gene Expression Levels. *Front. Microbiol.* **2016**, *7* (AUG), 1–14. https://doi.org/10.3389/fmicb.2016.01313.

(168) Pérez-López, M.; García-Contreras, R.; Soto-Hernández, M.; Rodríguez-Zavala, J. S.; Martínez-Vázquez, M.; Prado-Galbarro, F. J.; Castillo-Juárez, I. Antiquorum Sensing Activity of Seed Oils from Oleaginous Plants and Protective Effect during Challenge with Chromobacterium Violaceum. *J. Med. Food* **2018**, *21* (4), 356–363. https://doi.org/10.1089/jmf.2017.0080.

(169) Ravichandran, V.; Zhong, L.; Wang, H.; Yu, G.; Zhang, Y.; Li, A. Virtual Screening and Biomolecular Interactions of CviR-Based Quorum Sensing Inhibitors against Chromobacterium Violaceum. *Front. Cell. Infect. Microbiol.* **2018**, *8* (SEP), 1–13. https://doi.org/10.3389/fcimb.2018.00292.

(170) Qais, F. A.; Khan, M. S.; Ahmad, I. Broad-Spectrum Quorum Sensing and Biofilm Inhibition by Green Tea against Gram-Negative Pathogenic Bacteria: Deciphering the Role of Phytocompounds through Molecular Modelling. *Microb. Pathog.* **2019**, *126* (September 2018), 379–392. https://doi.org/10.1016/j.micpath.2018.11.030.

(171) Bodede, O.; Shaik, S.; Chenia, H.; Singh, P.; Moodley, R. Quorum Sensing Inhibitory Potential and in Silico Molecular Docking of Flavonoids and Novel Terpenoids from Senegalia Nigrescens. *J. Ethnopharmacol.* **2018**, *216* (January), 134–146. https://doi.org/10.1016/j.jep.2018.01.031.

(172) Reina, J. C.; Pérez-Victoria, I.; Martín, J.; Llamas, I. A Quorum-Sensing Inhibitor Strain of Vibrio Alginolyticus Blocks Qs-Controlled Phenotypes in Chromobacterium Violaceum and Pseudomonas Aeruginosa. *Mar. Drugs* **2019**, *17* (9). https://doi.org/10.3390/md17090494.

(173) Ohta, T.; Fukumoto, A.; Iizaka, Y.; Kato, F.; Koyama, Y.; Anzai, Y. Quorum Sensing Inhibitors against Chromobacterium Violaceum CV026 Derived from an Actinomycete Metabolite Library. *Biol. Pharm. Bull.* **2020**, *43* (1), 179–183. https://doi.org/10.1248/bpb.b19-00564.

(174) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47* (D1), D1102–D1109. https://doi.org/10.1093/nar/gky1033.

(175) Sander, T.; Freyss, J.; Von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program for Chemistry Aware Data Visualization and Analysis. *J. Chem. Inf. Model.* **2015**, *55* (2), 460–473. https://doi.org/10.1021/ci500588j.

(176) Sterling, T.; Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.

(177) Ortuso, F.; Bagetta, D.; Maruca, A.; Talarico, C.; Bolognesi, M. L.; Haider, N.; Borges, F.; Bryant, S.; Langer, T.; Senderowitz, H.; Alcaro, S. The Mu.Ta.Lig. Chemotheca: A Community-Populated Molecular Database for Multi-Target Ligands Identification and Compound-Repurposing. *Front. Chem.* **2018**, *6* (APR), 1–6. https://doi.org/10.3389/fchem.2018.00130.

(178) Atovaquone - DrugBank https://www.drugbank.ca/drugs/DB01117 (accessed Jul 21, 2020).

(179) Fenton, C.; Scott, L. J. Risperidone: A Review of Its Use in the Treatment of Bipolar Mania. *CNS Drugs* **2005**, *19* (5), 429–444. https://doi.org/10.2165/00023210-200519050-00005.

(180) Mebendazole - DrugBank https://www.drugbank.ca/drugs/DB00643 (accessed Jul 21, 2020).

(181) Bojar, R. A.; Cunliffe, W. J.; Holland, K. T. The Short-term Treatment of Acne Vulgaris with Benzoyl Peroxide: Effects on the Surface and Follicular Cutaneous Microflora. *Br. J. Dermatol.* **1995**, *132* (2), 204–208. https://doi.org/10.1111/j.1365-2133.1995.tb05014.x.

(182) Cilostazol - DrugBank https://www.drugbank.ca/drugs/DB01166 (accessed Jul 21, 2020).

(183) Mazìeres, B. Topical Ketoprofen Patch. *Drugs R D* **2005**, *6* (6), 337–344. https://doi.org/10.2165/00126839-200506060-00003.

(184) Permethrin - DrugBank https://www.drugbank.ca/drugs/DB04930 (accessed Sep 8, 2020).

(185) Gielen, W.; Cleophas, T. J.; Agrawal, R. Nebivolol: A Review of Its Clinical and Pharmacological Characteristics. *Int. J. Clin. Pharmacol. Ther.* **2006**, *44* (8), 344–357. https://doi.org/10.5414/CPP44344.

(186) Diamant, Z.; Sampson, A. P. Montelukast. *J. Drug Eval. Respir. Med.* **2002**, *1* (2), 53–88. https://doi.org/10.2165/00148581-200204020-00005.

(187) Diaz, G. A.; Krivitzky, L. S.; Mokhtarani, M.; Rhead, W.; Bartley, J.; Feigenbaum, A.; Longo, N.; Berquist, W.; Berry, S. A.; Gallagher, R.; Lichter-Konecki, U.; Bartholomew, D.; Harding, C. O.; Cederbaum, S.; Mccandless, S. E.; Smith, W.; Vockley, G.; Bart, S. A.; Korson, M. S.; Kronn, D.; Zori, R.; Merritt, J. L.; Nagamani, S. C. S.; Mauney, J.; Lemons, C.; Dickinson, K.; Moors, T. L.; Coakley, D. F.; Scharschmidt, B. F.; Lee, B. Ammonia Control and Neurocognitive Outcome among Urea Cycle Disorder Patients Treated with Glycerol Phenylbutyrate. *Hepatology* **2013**, *57* (6), 2171–2179. https://doi.org/10.1002/hep.26058.

(188) Vitamin K1 - Drugbank https://www.drugbank.ca/drugs/DB01022 (accessed Jul 20, 2020).

(189) Cazzola, M.; Testi, R.; Matera, M. G. Clinical Pharmacokinetics of Salmeterol. *Clin. Pharmacokinet.* **2002**, *41* (1), 19–30. https://doi.org/10.2165/00003088-

200241010-00003.

(190) Davis, R.; Whittington, R.; Bryson, H. M. Nefazodone. A Review of Its Pharmacology and Clinical Efficacy in the Management of Heart Failure. *Drugs* **1997**, *53* (4), 608–636.

(191) Grant, S. M.; Goa, K. L. Iloprost: A Review of Its Pharmacodynamic and Pharmacokinetic Properties, and Therapeutic Potential in Peripheral Vascular Disease, Myocardial Ischaemia and Extracorporeal Circulation Procedures. *Drugs* **1992**, *43* (6), 889–924. https://doi.org/10.2165/00003495-199243060-00008.

(192) Echizen, H.; Ishizaki, T. Clinical Pharmacokinetics of Famotidine. *Drug Metab. Dispos.* **1991**, *21* (3), 178–194.

(193) Quianzon, C. C. L.; Cheikh, I. E. History of Current Non-Insulin Medications for Diabetes Mellitus. **2012**, *1*, 2–5.

(194) Cefazolin - DrugBank https://www.drugbank.ca/drugs/DB01327 (accessed Jul 21, 2020).

(195) Cidofovir - DrugBank https://www.drugbank.ca/drugs/DB00369 (accessed Jul 21, 2020).

(196) Nizatidine - DrugBank https://www.drugbank.ca/drugs/DB00585 (accessed Jul 21, 2020).

(197) Pimozide - DrugBank https://www.drugbank.ca/drugs/DB01100 (accessed Jul 21, 2020).

(198) Urichuk, L.; Prior, T. I.; Dursun, S.; Baker, G. Metabolism of Atypical Antipsychotics : Involvement of Cytochrome P450 Enzymes and Relevance for Drug-Drug Interactions. *Curr. Drug Metab.* **2008**, *9* (5), 410–418. https://doi.org/10.2174/138920008784746373.

(199) Takasu, T.; Ukai, M.; Sato, S.; Matsui, T.; Nagase, I.; Maruyama, T.; Sasamata, M.; Miyata, K.; Uchida, H. Effect of (R) -2- (2-Aminothiazol-4-Yl) -4 J - { 2- [(2-Hydroxy-2- Phenylethyl) Amino] Ethyl } Acetanilide (YM178), a Novel Selective □ 3 -Adrenoceptor Agonist , on Bladder Function. **2007**, *321* (2), 642–647. https://doi.org/10.1124/jpet.106.115840.These.

(200) Sulfasalazine - Drugbank https://go.drugbank.com/drugs/DB00795 (accessed Oct 21, 2020).

(201) Strupczewski, J. T.; Bordeau, K. J.; Chiang, Y.; Glamkowski, E. J.; Conway, P. G.; Corbett, R.; Hartman, H. B.; Szewczak, M. R.; Wilmot, C. A.; Helsley, G. C. 3-[[(Aryloxy)Alkyl]Piperidinyl]-1,2-Benzisoxazoles as D2/5-HT2 Antagonists with Potential Atypical Antipsychotic Activity: Antipsychotic Profile of Iloperidone (HP 873). *J. Med. Chem.* **1995**, *38* (7), 1119–1131. https://doi.org/10.1021/jm00007a009.

(202) Frisch, M. J.; Trucks, G. .; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M. Gaussian 09, Revision A.02. Gaussian, Inc: Wallingford CT 2016.

(203) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095. https://doi.org/10.1021/ct400341p.