UNIVERSITÀ DEGLI STUDI DI SALERNO

DIFARMA

UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Farmacia

PhD Program

in **Drug Discovery and Development**

XXXIV Cycle — Academic Year 2021/2022

PhD Thesis in

High-throughput virtual screening by molecular dynamics assisted molecular docking.

Candidate

Anna Maria Nardiello

Supervisor

Prof. Dr. *Stefano Piotto*

Ph.D. Program Coordinator: Prof. Dr. *Gianluca Sbardella*

ABSTRACT

Nardiello Anna Maria, *High-throughput virtual screening by molecular dynamics assisted molecular docking*, Ph.D. program in Drug Discovery and Development, 2022, University of Salerno.

Molecular docking is a widely used Structure-Based method for drug discovery. Over the years, molecular docking methodologies have evolved to address several limitations, such as high errors in predicting pose and binding energy and approximations in the treatment of solvent molecules. These limitations were the starting point for my Ph.D. project, which involved using molecular dynamics (MD) and artificial neural networks to overcome these limitations. During the first year, I have addressed the problem of flexibility and high error in the calculation of the binding energy. From a thermodynamic point of view, water molecules can contribute to the formation of receptor-ligand complexes. During the second year, I evaluated the entropic contribution of water molecules in the binding site, developing a new docking protocol. During the third year, exploiting the studies started earlier on the entropy of water molecules in the active site, I developed a tool for binding site prediction. Moreover, I applied and validated the created protocol and other computational techniques to study the mechanism of action of the Sphingomyelin synthase system.

KEYWORDS: Molecular Docking, Virtual screening, Genetic algorithms, molecular dynamics.

PREFACE

My three-year Ph.D. course in Drug Discovery and Development at the University of Salerno started in November 2018 under the supervision of Prof. Dr. Stefano Piotto.

Molecular docking is a widely used structure-based method for drug discovery. Over the years, molecular docking methodologies have evolved to overcome several limitations, including errors in pose prediction and binding energy evaluation, and approximations in the treatment of solvent molecules. The most widely used docking programs simulate the flexibility of the protein in two ways: 1) by building different rotamers of the backbone side chains, or 2) by using different receptor conformations, obtained from NMR (Nuclear Magnetic Resonance) experiments. However, this is not sufficient to predict the conformational changes undergone by the receptor during the binding process. Another limitation of docking is that the solvent is not considered explicit. From a thermodynamic point of view, the release of water molecules during the binding event contributes to the change in binding free energy of the system by affecting both entropy and enthalpy. In all docking software, only the enthalpy change is considered explicitly, while the entropy change given by the movement of water molecules is not included. These limitations were the starting point for my Ph.D. project, which used molecular dynamics (MD) and genetic algorithms (GA) to create a new mathematical model for calculating the free energy of binding (NewΔG). To develop the new predictive model, I evaluated the error of two of the most widely used docking software (Vina, AutoDock) in predicting the binding pose and estimating the binding free energy. The binding pose geometry is a key point since docking algorithms rely on scoring values to determine the binding affinity, but if the geometry is wrong, the scoring is also wrong. Starting from a dataset of 300 receptor-ligand structures whose dissociation constant (K_D) values are known in the scientific literature, I predicted the geometry of the ligand pose via redocking and evaluated the correctness of the

prediction by calculating the RMSD (Root mean square deviation) between the experimental and predicted pose. I then calculated the error between the predicted and experimental binding energy values. I observed that the Vina algorithm is better than AutoDock in predicting both pose and binding energy in both cases. The values obtained from docking correspond only to the enthalpy contribution to the free energy change. To consider the entropic component, related to the movement of water molecules, I introduced a new parameter that determines the variation of the hydrogen bond energy over time. These values, along with the docking results, were used as descriptors to generate a model for predicting binding energy using GAs. Validation of the model on a set of 100 receptor-ligand crystallographic structures revealed a reduction of the average error in binding energy predictions of about 1.3 kcal/mol. During the last year, in collaboration with the University of Balearic Islands, where I spent my time abroad under the supervision of Prof. Pablo Escribà, I applied my predictive model to docking techniques and used other computational methods to study the possible mechanism of action of LP561, a drug currently in phase IIB clinical trials. LP561 is a 2-hydroxyoleic acid (2OHOA) used for the treatment of glioma. In diseased patients treated with 2OHOA, increased sphingomyelin levels with membrane stiffening have been observed. Sphingomyelin is synthesized by the enzyme Sphingomyelin Synthase (SMS), starting from phosphatidylcholine (PC) and Ceramide (Cer) and forming Sphingomyelin (SM) and diacylglycerol (DAG). Therefore, we decided to investigate this enzyme as a potential target of LP561. From the obtained results, it is hypothesized that LP561 is not a direct substrate of SMS, but most likely, given its nature of fatty acid could be incorporated into one of the two substrates of SMS and thus be indirectly responsible for the increased production of SM.

List of publications related to the scientific activity performed during the three years Ph.D. course in Drug Discovery and Development

Papers:

1) Mishra K, Péter M, **Nardiello AM**, Keller G, Llado V, Fernandez-Garcia P, Kahlert UD, Baraschd D, Saada A, Török Z, Balogh G, Escriba PV, Piotto S, Kakhlon O. Multifaceted analyses of isolated mitochondria establish the anticancer drug 2-hydroxyoleic acid as an inhibitor of substrate oxidation and an activator of complex IV-dependent state 3 respiration, Cells. 2022 (in press)
2) Sessa L, **Nardiello AM**, Santoro J, Concilio S, Piotto S. Hydroxylated fatty acids: The role of the sphingomyelin synthase and the origin of selectivity. Membranes. 2021;11(10).
3) Sessa L, Concilio S, Di Martino M, **Nardiello AM**, Miele Y, Rossi F, et al. A selective Nile Red based solvatochromic probe: A study of fluorescence in LUVs and GUVs model membranes. Dyes and Pigments. 2021;196.
4) Sessa L, **Nardiello AM**, Di Martino M, Marrafino F, Iannelli P. Molecular Dynamics Simulation of Antimicrobial Permeable PVC-Based Films. Lecture Notes in Bioengineering2020. p. 111-9.
5) Sarno M, Ponticorvo E, Piotto S, **Nardiello AM**, De Pasquale S, Funicello N. AuAg/ZnO nanocatalyst for CO2valorization and H2and CO electrochemical production. Journal of CO2 Utilization. 2020;39.
6) Piotto S, **Nardiello AM**, Di Biasi L, Sessa L. Encoding Materials Dynamics for Machine Learning Applications. Lecture Notes in Bioengineering2020. p. 128-36.
7) **Nardiello AM**, Piotto S, Di Biasi L, Sessa L. Pseudo-semantic Approach to Study Model Membranes. Lecture Notes in Bioengineering2020. p. 120-7.
8) Concilio S, Di Martino M, **Nardiello AM**, Panunzi B, Sessa L, Miele Y, et al. A flavone-based solvatochromic probe with a low expected perturbation impact on the membrane physical state. Molecules. 2020;25(15).
9) Piotto S, Sessa L, Piotto A, **Nardiello AM**, Concilio S. Plausible emergence of autocatalytic cycles under prebiotic conditions. Life. 2019;9(2).

Conference proceedings:

Nardiello A.M.; Presented at the BIONAM 2019, 3rd International Conference on Bio and Nanomaterials, Genoa (IT). MSC cruise, Mediterranean Sea, September 29 – October 3, 2019.

Table of contents

PART I:

New docking protocol for the free energy of binding estimation

INTRODUCTION

1. Introduction to molecular docking

The technique of molecular docking, originally introduced by Kuntz et al. [1], has developed increasingly in recent decades due to the growth of experimental data and the need to discover new drugs. Progress in this field has also been facilitated by the development of increasingly powerful computers and easy access to small molecule and protein databases.

Docking aims to predict how a small molecule (ligand) binds at the binding site of a macromolecular target (protein or nucleic acid). Docking, therefore, seeks to understand and predict the molecular and structural recognition between the two partners by predicting the structure of the [P+L] = [PL] complex under equilibrium conditions (Figure 1), and estimating the binding affinity.

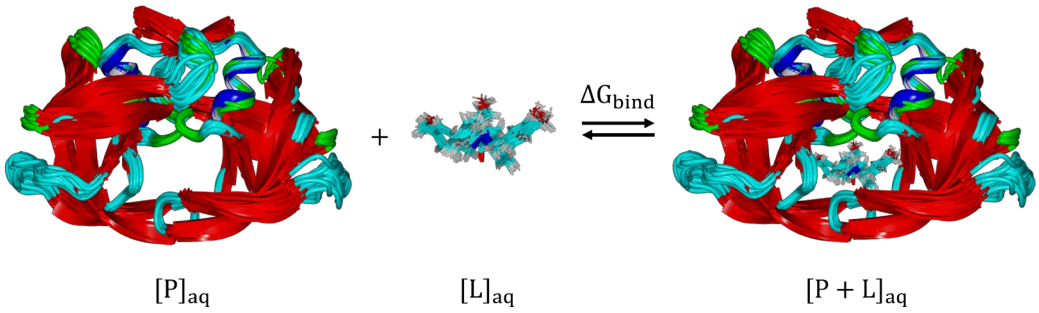

Figure 1 - Docking Scheme. The figure illustrates the binding of inhibitor Dmp323 to HIV protease and is based on solution structures (PDB code: 1BVE)

In many drug discovery applications, particularly in virtual screening, the docking is of the protein-ligand type. Only recently, docking has also been applied to predict the binding mode between two macromolecules, for example, between protein and protein or between protein and nucleic acids [2]. In this work, the focus is on protein-ligand docking.

Since its beginnings in the 1960s, molecular docking, along with developments in physics, chemistry, biochemistry and computer science, has become a powerful tool and an essential

technique in the discovery of new molecules, and many commercial drugs have been designed using computer-aided drug design methods (CADD) [3].

Over the past 15 years, interest in this technique has grown exponentially, as evidenced by the increase in the number of publications associated with molecular docking.

Figure 2 shows the number of publications since 1990 indexed in the PubMed search engine [4] associated with the keywords ‘docking' or 'dock' in the title or abstract.

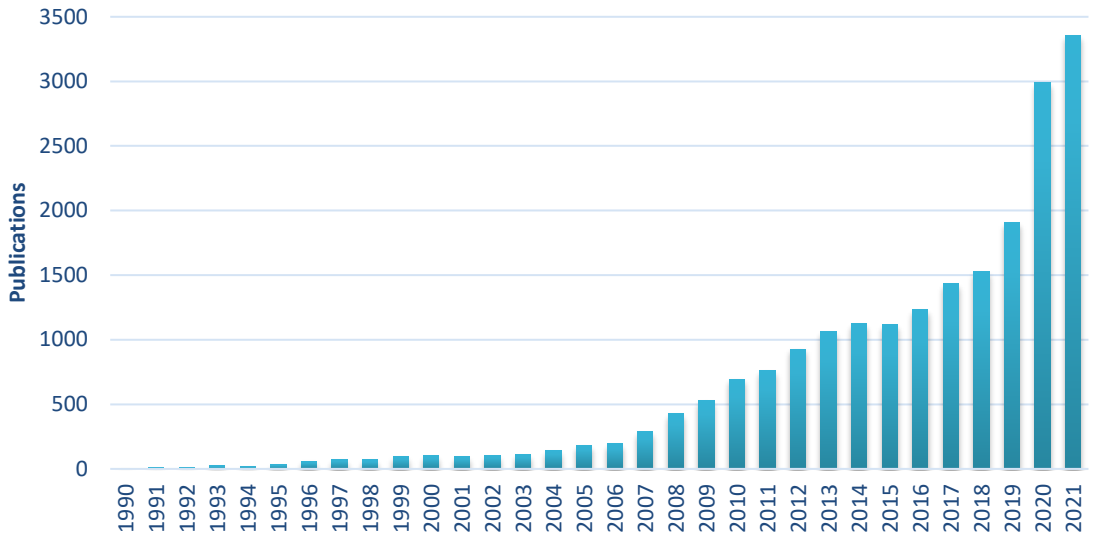

Figure 2 - Increase in the number of articles, from 1990 to 2021, retrieved from the PubMed Central (PMC)-NCBI database using 'Docking' or 'Dock' as keywords in the abstract or title

Molecular docking has a wide range of uses and applications in drug discovery, including structure-activity studies, lead optimisation and the search for potential leads through virtual screening.

2. Molecular Docking Theory

The underlying method of most docking algorithms is molecular mechanics. It describes a polyatomic system based on concepts from classical physics. Experimental parameters such as

charges, torsional and geometric angles are used to reduce the difference between experimental data and the predictions of molecular mechanics [5]. Due to the weaknesses and limitations of experimental parameters, mathematical equations can be parameterised based on semi-empirical and ab-initio quantum theoretical calculations. The set of equations describing biological systems is called molecular force fields. Since force fields use different approximations and simplifications, the description of the system can be inaccurate.

Most force fields are based on five terms [6]:

- potential energy
- torsional terms
- bond geometry
- electrostatic terms
- Lennard-Jones potential.

Examples of relevant force fields are AMBER, GROMOS, MMFF94, CHARMM and UFF. With the development of force fields, molecular modeling has focused on simulating binding processes between protein and ligand.

Two general models have been developed to simulate the interaction between two molecules. The first is the rigid-body approach (lock and key model), closely related to the classical model of Emil Fischer [7]. In this model, the ligand and the receptor are considered as two independent and rigid bodies that interact according to their shape and volume. The second approach is flexible docking (induced-fit model). This model assumes a mutual recognition between protein and ligand with a consequent adaptation of each part to the other [8]. Figure 3 outlines the two described approaches.

Figure 3 - Schematic representation of the two main approaches for molecular docking: rigid body and induced-fit.

Molecular docking has two main purposes. The first is to predict and correctly identify the binding mode of a ligand in the active site of a protein. The second is to correctly classify the poses of complexes according to their binding affinities [9, 10].

All docking protocols have two essential components: a positioning algorithm and a classification system or scoring function. However simple the process may seem, both parts bring complex problems [11, 12].

Positioning requires an exhaustive exploration of the accessible conformational space and binding orientations within the active site to extensively map the interactions between the residues and the ligand. This process needs to be accurate while maintaining a reasonable speed. Furthermore, the ability to correctly score and rank the poses generated for a ligand presents an even greater challenge as many scoring functions fail to predict binding affinity accurately and often report a score that may or may not be consistent with experimentally measured binding affinities [13].

The first step in any molecular docking protocol is to identify the recognition region between the two interacting species. This region may be located at the molecular surface of the two species or, as in the case of small organic molecules and macromolecular systems, may be located in special cavities called binding pockets. These are formed by the folding of the surface of the macromolecular structure, within which the ligand can be housed and perform its activity.

The identification and mapping a binding site can reveal key elements in protein-ligand binding [14]. Such knowledge is indispensable for docking and for the design of new drugs because, in most cases, receptor-drug interactions are specific [15]. A comprehensive suite of binding cavity detection methods has been developed to address these issues in docking and virtual screening simulations. Among these, the creation of a new tool for identifying potential binding sites by exploiting studies on the movement of water molecules solvating the system will be described in section 10.4.

Successful docking is largely dependent on the quality of information regarding the architecture of the active site, as it is the size and shape of the active site itself that dictates the three-dimensional geometry of the ligands that can bind there. Thus, a clear definition of the surface of a binding pocket, together with the identification of protein-ligand interaction sites, provides a set of features for ligand orientation.

The shape and size of the binding pockets are also potentially subject to significant variations caused by the rotation of amino acid side chains, backbone movements, loop movements and/or conformational changes induced by the ligand [16]. The geometry in which the ligand is placed in the cavity is a function of both its steric requirements and the nature of the electrostatic interactions it will establish in the cavity.

Proteins are known to exploit their inherent conformational flexibility to perform a wide range of biochemical processes [17]. In many cases, subtle movements in the domains, flexibility in the main chain of the protein, or reorientation of the side chains, change the shape and size of the pocket as it binds to the ligand [18]. These changes include hinge movements of entire domains, small rearrangements of side chains in binding pocket residues [19], and even structural transitions involving the opening/closing of otherwise rigid domains of the protein (e.g. the opening of TM6 in GCPRs) [20]. Currently, most docking approaches treat ligands as

flexible, but incorporating protein flexibility into the docking protocol remains a challenging task. An in-depth analysis of side-chain flexibility can provide valuable insights to improve docking performance and optimise protein-ligand interactions.

Each docking algorithm can be simply described as a combination of a search algorithm and a score function.

3. Search algorithms and pose prediction

The search for a pose in a docking protocol is of utmost importance. The success of an algorithm in predicting a ligand-binding position is normally measured in terms of the Root Mean Square Deviation (RMSD) between the experimentally observed ligand positions and those predicted by the algorithm. The system's flexibility is a major challenge in searching for the correct pose. The number of degrees of freedom included in the conformational investigation is a central aspect that determines the efficiency of the search [21].

The ligand, macromolecular receptor, and solvent molecules are present in a real biological system. Solvent molecules are generally excluded from the simulated system due to the enormous number of freedom degrees associated with them. In special cases, they are modelled implicitly in the scoring functions. However, the part of the system consisting only of ligand and receptor also has several degrees of freedom that is computationally untreatable, so the dimensionality of the problem must be reduced by applying various approximations.

There are several levels of approximation. The simplest is the rigid-body approximation, very popular in early approaches to docking (and still widely applied in the field of protein-protein docking), which treats both the ligand and the receptor as rigid and explores only the 6 degrees of translational and rotational freedom, thus excluding any kind of flexibility.

A more common approach is to model the flexibility of the ligand while assuming a rigid receptor [21], thus considering only the conformational space of the ligand. There are three general categories of algorithms designed to deal with ligand flexibility:

- Systematic search methods
- Random or stochastic methods
- Simulation methods.

3.1. Systematic search algorithms

Systematic search algorithms support small variations in structural parameters, progressively changing the conformation of ligands. Systematic search algorithms attempt to explore all the states of freedom of a molecule based on bond rotations, angles, and increments' size [22]. Due to a large number of conformations, systematic searches have to deal with the problem of combinatorial explosion [23].

Systematic search algorithms can be further divided into three main types:

- conformational research methods
- fragmentation methods
- database methods.

Conformational search methods can be seen as the solution to the problem of flexible ligand docking. All rotatable bonds in the ligand are systematically rotated 360° using a fixed increment until all possible combinations have been generated and evaluated. A major pitfall in this type of method is that the number of generated structures increases exponentially with the number of rotatable bonds, a phenomenon known as combinatorial explosion [23]. Therefore,

the application of this type of method is very limited. Generally, various constraints and restrictions on the ligand must be employed to reduce the dimensionality of the problem.

Fragmentation is one of the most commonly used approaches to introduce ligand flexibility into molecular docking. Fragmentation methods incrementally grow ligands in the active site, either by inserting the various fragments into the active site and covalently attaching them to recreate the initial ligand (place-and-join approach) or by dividing the ligand into a rigid central fragment that is inserted into the active site and flexible regions that are added later (incremental approach).

Database methods address the combinatorial explosion problem by using libraries of pre-generated conformations (conformational ensembles).

3.2. Stochastic or random search methods

Stochastic or random search methods are based on random changes to a single ligand or a population of ligands that are evaluated with a predefined probability function. Derived from the probability criterion, favourable changes are accepted. The algorithm generates sets of molecular conformations and populates a wide range of the energy landscape. This strategy avoids trapping the final solution in a local energy minimum and increases the probability of finding a global minimum. Since the algorithm promotes a wide coverage of the energy landscape, the computational cost associated with this procedure is a major limitation.

There are three types of methods based on random algorithms:

- Monte Carlo methods (MC)
- Genetic Algorithm methods (GA)
- Tabu Search methods.

In Monte Carlo methods, the acceptability criterion for a newly obtained pose is based on a Boltzmann probability function. MC methods have the advantage of using a simple energy function that does not require any kind of derived information [24]. In addition, they are efficient in overcoming energy barriers, thus enabling more comprehensive searches of the conformation space.

Genetic algorithms (GA) apply ideas derived from genetics and the theory of biological evolution to docking. In contrast to standard MC methods, GAs start with an initial population of different ligand conformations with respect to the protein.

Each conformation is defined by a set of state variables (genes) that describe aspects such as translation, orientation, and conformation of the ligand in relation to the protein. The complete set of ligand state variables is referred to as the genotype, while the atomic coordinates refer to the phenotype. Genetic operators (mutations, crosses, and migrations) are applied to the population to sample the conformational space until a final population is reached that optimises a predefined fitness function.

The Lamarckian genetic algorithm, for example, is implemented in AUTODOCK, which switches from genotypic space to phenotypic space. Mutation and crossover occur in the genotypic space while the phenotypic space is decided by the energy function and optimised. From energy minimization, phenotypic alterations are mapped back to the genes through the change in ligand state variables.

Tabu search methods operate by imposing restrictions that prevent the search from revisiting already explored areas of the conformational space, promoting the analysis of new regions. This is achieved through a list that stores previously visited solutions. Calculation of the RMSD of a new conformation against all previously recorded conformations of the ligand determines whether the new conformation is accepted.

3.3. Simulation methods

Simulation methods take a rather different approach to the docking problem and are based on calculating solutions to Newton's equations of motion. There are two main types: molecular dynamics (MD) and energy-minimisation methods.

Molecular dynamics methods are a powerful and versatile tool in studying a wide range of applications [25]. However, despite the increasing popularity of these methods in docking, several pitfalls are well known. Difficulties in crossing high-energy barriers and the problem of sampling the conformational space within an acceptable simulation period are the main drawbacks for the application of MD in protein and ligand docking. Some strategies have been found to compensate for these limitations, such as using very high temperatures.

Energy minimisation methods include direct searches (simplex), gradient methods (steepest descend), conjugate gradient methods (Fletcher-Reeves), secondary derivative methods (Newton-Raphson), and least squares methods (Marquardt), which are rarely used as a stand-alone search technique in docking because only local minima can be achieved. However, many of the other docking algorithms described above commonly use energy-minimisation methods as a complement.

Thus, molecular dynamics simulation can localise ligands within local minima. Complementing other methods (such as simulated annealing) followed by molecular dynamics simulation can provide better results. In contrast to MD simulation, energy-minimisation methods are hardly ever used as the sole search technique.

4. Scoring Function

Scoring functions are fast and approximate mathematical models used to evaluate the binding affinity (usually by measuring non-covalent interactions) between the protein and the ligand. A perfect scoring function would be able to predict the free energy of binding of the protein-ligand complex and at the same time be fast enough to allow its application to virtual screening.

To accurately calculate the free energy of binding, many physical interactions (especially those involving the solvent, including entropic effects) should be included; but this is unrealistic due to the complexity of the algorithm and the need for large calculations. As a result, scoring functions incorporate several simplifications to reduce the complexity of calculation at the cost of accuracy [26]. The lack of an appropriate scoring function, both speed and accuracy, is the main bottleneck in the molecular docking technique [27]. The scoring functions normally used in protein and ligand docking can be divided into three main classes:

- Force-field based scoring
- Empirical scoring
- Knowledge based scoring

4.1. Force-field based scoring

Force-field based scoring methods generally use a molecular mechanical force field. Standard force fields quantify the sum of the interaction energy between the receptor and the ligand and the internal energy of the ligand. The energies are normally evaluated by a combination of van der Waals energy and electrostatic energy terms.

The Lennard-Jones potential is used to describe the van der Waals energy term, while the electrostatic term is given by a Coulombian formula with a distance-dependent dielectric function that decreases the contribution of charge-charge interactions [27].

The popularity of these methods in virtual screening is a consequence of their simplicity. Although faster and simpler, these functions are not ideal for simulating biomolecular interactions, as they were developed to calculate the enthalpy of binding in the gas phase. Traditional limitations of scoring functions include the absence of solvation and entropic terms, and the inaccurate treatment of long-range effects involved in binding.

4.2. Empirical scoring

Empirical scoring functions are designed to reproduce experimental data and are based on the idea that binding energies can be approximated by a sum of several individuals and uncorrelated terms [28]. The rationale is that the free energy of binding of a non-covalent protein-ligand complex can be factored into a sum of localised and chemically intuitive interactions. Terms representing different contributions such as hydrogen bonds, hydrophobic interactions, entropic effects are normalised by weighting factors derived from regression analysis of data from well-characterised protein-ligand complexes. The binding affinity is estimated as a sum of interactions multiplied by the weighting factors and solved by an equation of the type:

$$\Delta G_{binding} \approx \sum \Delta Gi \; fi \; (rl, rp)$$

Where *fi* is a simple geometric function of the coordinates of the ligand (*rl*) and the receptor (*rp*).

The interest in empirical scoring functions stems mainly from the ease with which the various terms can be calculated, whereas the main disadvantage of these methods lies in their dependence on the experimental data set used in the parameterisation process (not versatile and not transferable).

4.3. Knowledge-based scoring

Knowledge-based scoring functions are based on rules derived from structural data analysis of known and well-characterised receptor-ligand interactions [29]. The exponential growth and availability of protein-ligand crystal structures is enabling the derivation and formulation of rule sets based on the frequencies of chemical interactions.

This type of scoring function attempts to capture knowledge about the protein-ligand binding that is implicitly stored in the protein database through statistical analysis of structural data. Potentials are obtained from the statistical analysis of the coupling frequencies of atoms observed in the crystal structures of protein-ligand complexes [30]. The accuracy of this scoring function depends on the quality of the experimental data, as it incorporates structural knowledge without considering inconsistencies in the experimental and structural data.

A significant advantage of knowledge-based scoring is its balance between performance and computation time. In addition, it can consider uncommon interactions such as sulphur-aromatic interactions [31].

Over the years, there have been attempts at improvement leading to 'hybrid' approaches, combining empirical data and knowledge-based potentials [32].

5. Most common docking software

There are several servers, suites, and programs available for molecular docking. Each tool uses different algorithms for pose generation, refinement, and calculation of receptor-ligand interactions. Table 1 presents a shortlist of the main docking programs, including their algorithms and general information.

Table 1 - Most commonly used software and algorithms for docking

Name	Search algorithm	Type
AUTODOCK4	Lamarckian genetic algorithm	Academic
DOCK	Shape matching	Academic
OEDOCKING	Shape matching	Academic
FLEKSY	Ensemble-based	Commercial
SWISSDOCK	Evolutionary optimization	Academic
GOLD	Genetic algorithm	Commercial
GLIDE	Hybrid	Commercial
VINA	Local optimization	Academic
RDOCK	Hybrid	Academic
LEDOCK	Simulated annealing	Academic
PLANTS	Ant colony optimization	Academic
HADDOCK	Hybrid	Academic
SURFLEX-DOCK	Shape matching	Commercial
MOE	Hybrid	Commercial
FLEXX	Shape matching	Commercial

6. Limitations of docking

As mentioned above, docking is a computational technique for predicting the interaction between a ligand and a protein. A molecule is placed inside the target protein's binding cavity, and the predicted pose is evaluated by a scoring function [32, 33]. The latter generates a score

for each predicted pose, and the resulting values are used to discriminate between different ligands. As docking is typically used to screen large libraries of ligands, the computational cost must be low.

To meet this requirement, several simplifications have been applied [34]. The reduction in flexibility is the first. While small molecules can be considered fully flexible, the same is not true for proteins. The most common docking software simulates protein flexibility by constructing rotamers of the side chains or using different receptor conformations from NMR experiments. The conformational changes undergone by the receptor during the binding process are therefore underestimated [35]. Another simplification relates to water, which is not explicitly considered. A change in the orientation of a single water molecule in the binding site not only has an effect on neighbouring waters but also extends to the surrounding hydration layers, affecting the entire hydrogen-bonding network [36]. This network is decisive for the calculation of the free energy variation of the system [37]. However, in all of the most commonly used approaches, only the enthalpy variation is calculated, but not the entropy contribution. The main limitations of this computational technique will be discussed in detail in the following paragraphs.

6.1. Receptor flexibility

Historical lock-and-key and induced-fit theories have given way to more modern theories that give greater weight to the problem of receptor flexibility [38].

The flexibility of the receptor and ligand is one of the main challenges in docking. A correct approach to test the behaviour of a protein-ligand complex is in a dynamic environment. New folding patterns have been discovered with the exponential growth of databases such as the

Protein Data Bank [39]. Docking methods with the flexible ligand generally give good results for about half of the studies to which they are applied [39, 40].

These success stories include systems in which the pose of the complex, which performs the biological function, is quite rigid and has not undergone major changes in structure upon binding to the ligand. However, many systems show significant movement upon ligand binding, and even small movements such as local rearrangements of side chains have an important effect on docking results.

The development of computational strategies that consider protein flexibility in the context of docking is still in its infancy, but several approaches have been devised that can at least partially introduce flexibility into the protein. These include some molecular dynamics and Monte Carlo methods [41, 42], rotamer libraries [17], grids of protein arrays [43], and Soft-receptor modelling [40, 44]. The basic principles previously described for molecular dynamics and Monte Carlo methods in the context of protein-ligand docking for flexible ligands are also applicable for receptor flexibility. In this case, however, the size of the problem and the research space is considerably increased.

Docking simulations with a fully flexible target are currently not feasible, due to the need to obtain a docked complex with a computation time of a few minutes. A few studies with fully flexible proteins have been reported in the literature, but they required several days of calculation [45].

Many methods have been devised to simplify the molecular simulation of the system, allowing the incorporation of the limited movement of the protein, while keeping the computational cost to a minimum. Methods based on rotamer libraries attempt to represent the protein's conformational space as a set of experimentally observed rotamers for each side chain [46, 47]. Focusing on the side chains neglects any real change in the protein's backbone; to

reasonable factor in protein flexibility, it is necessary to go beyond simple side-chain reorientation [48].

The use of multiple protein conformations of a single structure is considered an alternative strategy to address protein flexibility. Several approaches have explored this basic idea, but questions remain as to which is the best source of multiple protein structures (crystal structures, NMR, or calculations) and how to combine the information obtained from the different conformations [48]. Furthermore, some studies of ligand docking using an ensemble of protein structures have resulted in success rates worse than rigid docking itself [49].

6.2. Solvent inclusion

Water molecules play multiple roles in the structure and function of biological systems and often play a critical role in modulating protein-ligand interactions. The importance of water molecules in the binding site cannot, therefore, be underestimated [50]. The network of hydrogen bonds related to the water molecules in the binding site has an influence both in the evaluation of the structure-activity relationship [51], and in the optimisation of the ligand, taking into account that a higher binding affinity and a longer residence time can be achieved [52].

The problem with explicitly considering the solvent is that detailed information about the water in and around the binding site is unavailable. X-ray crystallography is the most common tool for determining the 3D structure, which can only provide partial information because the resolution and low electron density limit water detection. In addition, crystallisation conditions are typically far from biologically relevant, and co-crystallised ligand molecules may also influence the observed hydration network (in a different way to a binding ligand).

Considering water in the docking process is a complex task. A single water molecule has limited rotational freedom and the ability to form hydrogen bonds. For protein-ligand complexes, many water molecules are retained in the active site and contribute to the binding energy between protein and ligand, regardless of entropic considerations. Water can act as a bridge between the protein and the ligand and allow what would otherwise be unfavourable interactions between two chemically incompatible groups (e.g., between two bases). Water molecules can also modify the shape and microenvironment of the active site by associating closely with specific residues and thus presenting a different steric and electrostatic profile of the binding pocket than that presented by an anhydrous active site [53]. These various functional involvements of water define a number of important considerations that must be respected in quality docking experiments and rational design of high-affinity molecules. The surface areas accessible to water molecules, the hydrogen bonds involving water, the conservation and/or displacement of water, as well as the interaction energy of the molecules are just some of the factors that need to be considered in docking simulations. The reality is that state-of-the-art docking algorithms, and the associated scoring functions, do not adequately consider all the contributions of water molecules explicitly.

6.3. Consideration of system entropy

Entropic contributions from ligand binding cannot be underestimated, but are often undervalued and not quantified [54]. Entropy is not easy to calculate and for this reason it is a parameter that is often sacrificed in the calculation of free binding energy in favour of computational efficiency. Frequently, docking algorithms only consider the enthalpic contribution to the binding energy and not the entropic contribution of the solvent [55]. When

a ligand interacts with the protein's binding site, it must displace the water molecules that occupy it. Bulk water molecules have a greater possibility of establishing hydrogen bonds. Displacement of water molecules from the binding site results in an increase in the system's entropy [56].

Entropy, in statistical mechanics, is a measure of the disorder in a physical system. An increase in disorder is associated with an increase in entropy. The study of entropy is a way to obtain macroscopic information from microscopic configurations: to a certain macroscopic condition of equilibrium of the system (macrostate or thermodynamic state of the system, defined by precise values of quantities such as pressure and temperature) correspond different microscopic configurations (dynamic states or microstates, defined only if the position and velocity of all the molecules of the system are known).

Then we can define entropy according to Boltzmann's principle as:

$$S = k_B \ln \Omega$$

Where k_B is the Boltzmann constant, Ω is the number of microstates.

Since entropy is directly related to the number of microstates, in docking experiments it is not possible to truly estimate it because when analysing a pose, it is only a snapshot of the binding process.

7. Artificial intelligence tool: genetic algorithms.

Genetic algorithms are artificial intelligence tools inspired by Charles Darwin's theory of natural evolution. These algorithms reflect the process of natural selection in which the most suitable individuals are selected to reproduce the next generation's progeny. The process of natural selection begins with the selection of the fittest individuals in a population. They

produce offspring that inherit the characteristics of their parents and will be added to the next generation. If the parents are in better shape, their offspring will be better than the parents and have a better chance of surviving. This process continues to iterate and eventually a generation with the most suitable individuals will be found. This notion can be applied to various problems. A set of solutions for a problem are considered and the best set of them is automatically selected.

When we use artificial intelligence techniques, such as genetic algorithms, it is common to run into overfitting problems. Usually, a learning algorithm is trained using a defined set of examples (the training set), of which the target parameter (output) is already known. The learning algorithm (the learner) will reach a state where it will be able to predict outputs for an unknown dataset, i.e., it is assumed that the learning model will be able to generalise. However, especially when learning has been carried out for too long or when the training set is relatively small, the model may fit features that are specific to the training set only, without being able to generalise; therefore, in the presence of overfitting, performance (i.e., the ability to fit/predict) on the training data will be high, while performance on unknown data will be low.

AIM OF THE WORK

8. Aim of the work and thesis outline

One of the major aspirations of in-silico drug design is to accurately predict the binding affinity of a molecule to its target protein in order to obtain a potent and selective drug. The most widely used computational tool is molecular docking. Despite the progress made in this field, docking algorithms still show margins for improvement and are undergoing continuous development to maximise the accuracy of the predicted binding energies and to minimise computational error. The main problem, encountered over the years, is that even when structural information of the protein/ligand complex is available, parameters such as the effect of solvents, conformational changes of the protein and/or ligand and evaluation of the entropy and enthalpy of the system are still neglected.

Starting from these limitations, this PhD project aimed to develop a new mathematical model for predicting the binding energy more accurately than commonly used docking software. Focuses of the new mathematical model include the full flexibility of the protein and the explicit consideration of the solvent using molecular dynamics techniques.

Creating an accurate and fast docking protocol was not an easy challenge. Artificial intelligence tools were needed to optimise and create the new mathematical model for estimating the free energy of binding (ΔG).

Initially, receptor-ligand complexes were selected for which the experimental values of free binding energy and the three-dimensional crystallographic structure were known.

300 structures of receptor/ligand (R/L) complexes were downloaded from the RCSB database [57] and the corresponding experimental values obtained from the BindingDB database [58] (the term receptor will be used to refer to any protein target throughout this thesis). A study of the performance (in terms of RMSD and estimation of free binding energy) of two

of the most widely used algorithms in docking was carried out. Numerous studies on this topic are existing in the literature, but it was necessary to perform a new study with the objective of considering the error in the prediction of the energy correlated to the binding pose. In fact, it is well known that one of the main problems of docking is the scoring, since not always the best estimated pose is the one corresponding to the experimental pose.

Analyses of the movement of the water molecules solvating the system and the network of hydrogen bonds formed by them were carried out to include entropic considerations in the estimation of the free energy of binding.

The data from the water study and the output data from the docking experiments were used to create a mathematical model for the estimation of the free energy of binding, using genetic algorithms (GA). The model obtained was finally validated on a new dataset of complexes.

The model obtained was finally validated on a new dataset of R/L complexes.

Taking advantage of studies carried out on water molecules in the binding site, a tool that could predict the presence of possible binding pockets, called WaterScope, was developed. This plugin is based on the idea that water molecules occupying the binding site have a greater propensity to be displaced, so those occupying the binding site are the molecules with the lowest turnover of hydrogen bonds.

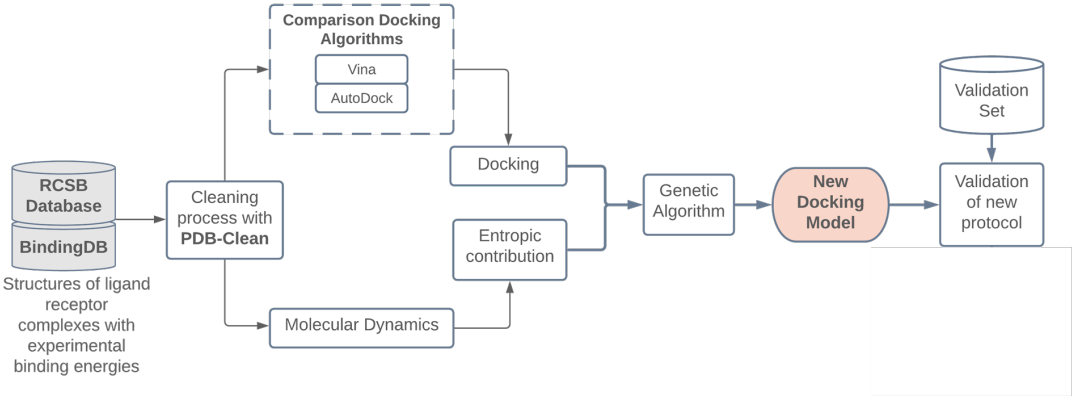

Figure 4 – Docking workflow performed in this job

MATERIALS AND METHODS

9. Dataset Selection

BindingDB [58] is a public accessible database containing more than 20,000 experimentally determined binding affinities of protein-ligand complexes. Data are extracted from the scientific literature, data collection focuses on proteins that are drug targets or candidates and for which structural data are present in the Protein Data Bank [57]. From this database, 300 receptor-ligand complexes were selected. The complexes were chosen in a heterogeneous manner, initially eliminating membrane proteins and those whose size exceeded 600 amino acids. The set was then subdivided into a training set of 200 structures, the remainder constituting the validation set. The training set was then further split into smaller sets to avoid overfitting when applying the genetic algorithms.

10. Automatic processing of PDB format: PDBClean

The PDB (Protein Data Bank) format was created in 1971, when the possibility of graphically displaying the 3D structures of proteins on a computer did not exist. This problem was overcome by noting down in a file all the spatial positions of the atoms making up the proteins, obtained by X-ray crystallography and NMR spectroscopy.

One of the main inconveniences of working with PDB files is that they contain not only information about atoms of interest, such as those of a protein or ligand, but also information about solvent molecules, the water of crystallisation, ions, metals, and other non-protein molecules, which are of no interest for most computational studies.

Figure 5 shows the number of times certain chemicals of minor interest for docking studies, appear in PDB files deposited in the RCSB database [57].

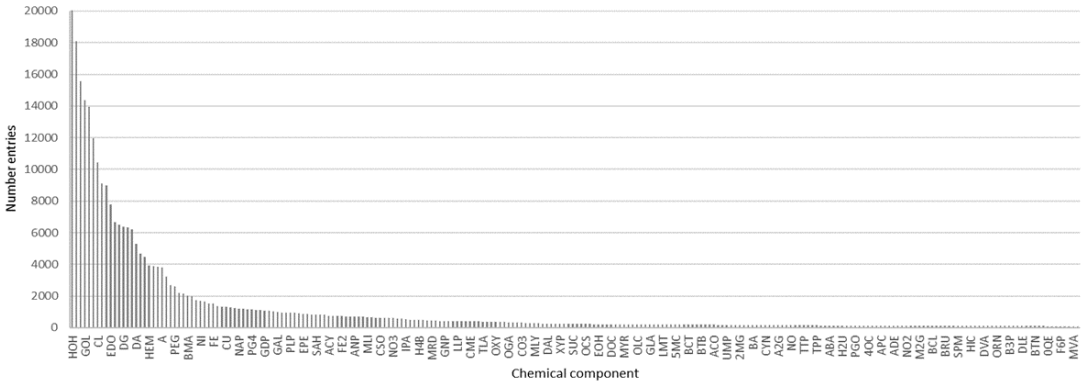

Figure 5 - Entries of the most frequent chemicals present in PDBs on RCSD Protein Data Bank

Since most of the analyses performed on PDBs require the removal of unnecessary components from the file, a plugin was developed for YASARA software that enables the automatic processing of such files.

PDB processing was performed by comparing the molecules present in the PDB file of interest with lists obtained from ligand-expo.rcsb, filtered to extract the molecules most commonly used in sample preparation processes for X-ray or NMR. The plugin recognises water molecules useful for interacting protein and ligand in the binding pocket. Coordinate metals are not eliminated, but recognised and considered part of the receptor, as well as any cofactors.

Figure 6 shows an example of a PDB (PDB ID: 1AVN) processed with the plugin and shows how the crystallisation water, the azide ion and the mercury atom have been eliminated, while the zinc atom has been preserved.

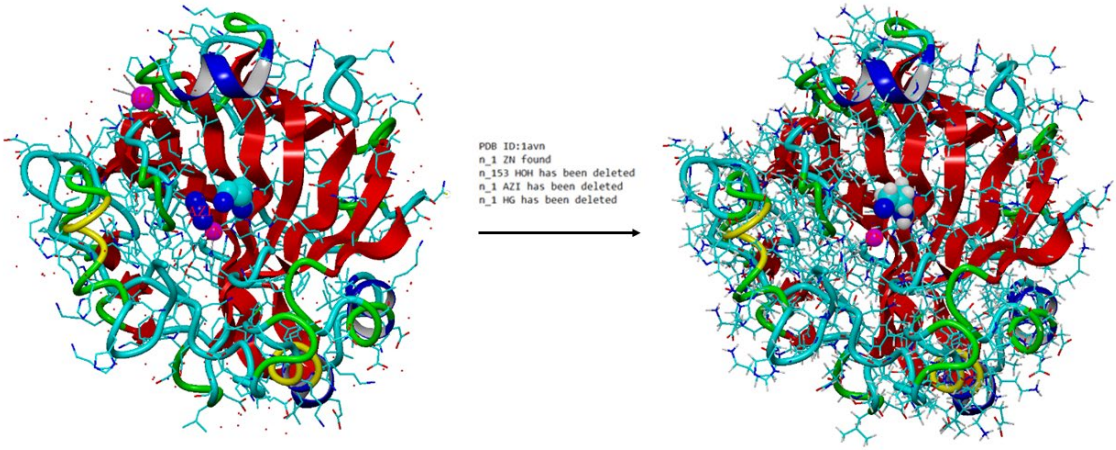

Figure 6 - Using the PDBClean plugin on human carbonic anhydrase II complexed with the histamine activator. The Zinc atom (fuchsia sphere) has been preserved. (PDB ID: 1AVN)

This plugin has enabled the rapid processing of a large quantity of PDBs. This step becomes essential for developing an automated and fast docking protocol.

11. Autodock and Vina Docking

Autodock and Vina are two of the most widely used docking algorithms implemented in the Yasara Structure software [59]. Docking experiments were performed starting with the PDB of the experimental complex previously processed with PDBClean, placing a 5Å simulation cell around the ligand.

For each best-pose resulting from docking, the accuracy of the pose prediction was assessed by calculating the RMSD (Root mean square deviation) between the algorithm best-pose and the experimental best-pose. As Autodock is an algorithm whose scoring function depends on the force field, AMBER15FB was chosen. For the poses that had an RMSD of less than 2Å, the binding energy prediction was also evaluated and compared to the experimental one by means of R^2 and the calculation of the mean error in Kcal/mol.

12. Molecular dynamics and entropic parameters

A molecular dynamics (MD) simulation was performed on each complex to solvate the entire system and to have full ligand and receptor flexibility using Yasara Structure software [59]. The force field is AMBER15FB under NPT conditions, with the Berensend thermostat. The applied cut-off is 8Å with Particle Mesh Ewald (PME). All systems were hydrated to a water density of 0.997 g/mL and neutralised with NaCl at a concentration of 0.9%. The ligand was removed so that the binding pocket was hydrated. A simulation of a 1ns was performed (Figure 7).

The entropic contribution of the water molecules occupying the binding site has been estimated by calculating parameters such as the variability of the network of hydrogen bonds forming between water molecules, their average energy, and the average number of water molecules in the binding site. These parameters will be described in detail in the following paragraphs.

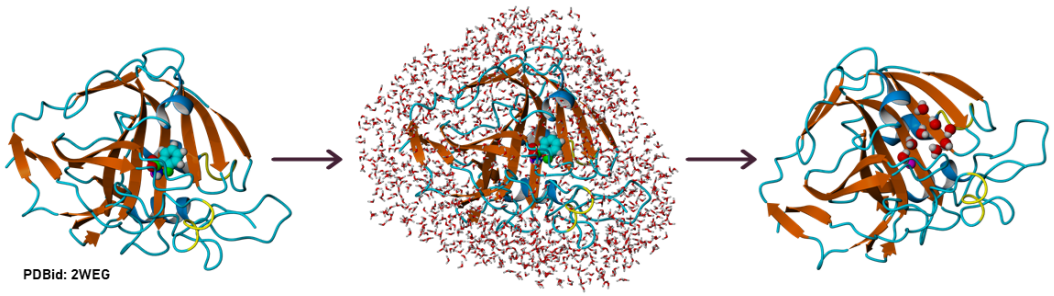

Figure 7 - Solvation process of the system. The ligand is removed and then solvated again to include the water molecules in the binding pocket.

13. Genetic function approximation (GFA)

Genetic algorithms are search algorithms that take their inspiration from natural and evolutionary genetics. The software used to create the model is Biovia Materials Studio [60]. In particular, the GFA (Genetic function approximation) approach was used in which, given a large number of raw inputs, the subset of terms that correlates best with the selected target response is found.

The selected target parameter is the experimental binding energy of the ligand-receptor complexes, and the other terms constitute the population. The population evolves to reach the chosen target. Selection, crossover, and mutation are then performed iteratively in succession. The procedure continues for a specified number of generations, until a convergence triggered by the lack of progress in the population scores.

The number of descriptors in the regression equation is set to 2, and the population and generation are set to 1000 and 1500 respectively. The resulting models are 10 and are evaluated on R^2, which is the fitness of the model.

14. ROC Curve and AUC analysis

Measuring the performance of the generated model is an essential task and can be done by calculating the ROC (Receiver Operating Characteristics) curve and its AUC (Area Under the Curve). It is one of the best evaluation metrics to check the performance of any classification model, it allows to verify how well the model can distinguish between classes. ROC curves are generally used to evaluate the performance of a classifier by considering the True Positive Rate (TPR) and False Positive Rate (FPR) for a set of data. The TPR, also known as sensitivity, is the percentage of target samples correctly classified as positive, while the FPR, also known as

(1-specificity), is the percentage of samples incorrectly classified as positive. The reliability of the scoring function was estimated using the area under the curve. An AUC value close to 1 indicates good selectivity, while a value below 0.5 indicates random selection. Typically, a ROC curve has an AUC baseline of 0.5 which demonstrates a uniformly distributed system.

RESULTS AND DISCUSSION

15. Discussion

15.1. Water Network variability

Water plays an extremely important role in all biological processes due to its peculiar physical and chemical properties. During the binding process, water molecules solvating the binding pocket and the ligand must be displaced to allow interaction between them (Figure 8).

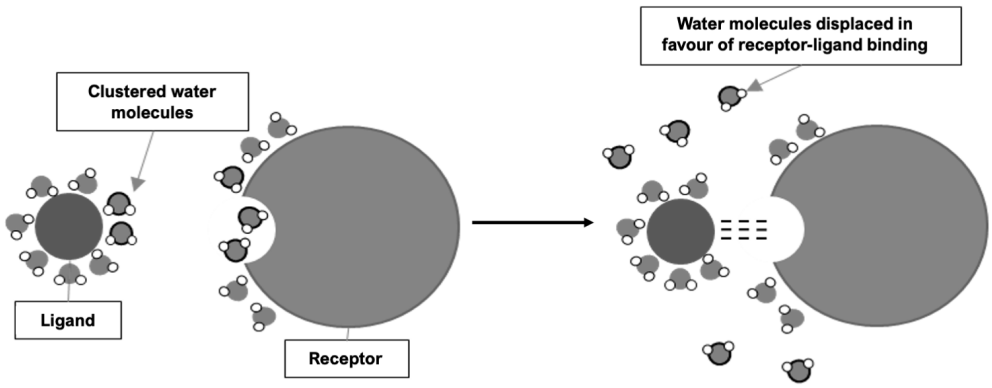

Figure 8 - Receptor and ligand desolvation to allow the formation of the complex

Water molecules have different propensities to move if confined in a bonding pocket or free to move in the bulk. The movement of water molecules must be considered to reflect this different propensity to displacement on the calculation of the system's entropy. Therefore, monitoring how much the water molecules present in the binding site changed their hydrogen bonding network was planned. The exchange of hydrogen bonds of each water molecule was monitored by counting the number of bonds at each step of the simulation.

When analysing the MD trajectories, the first problem encountered was the duration of these bonds. In fact, 90% of the monitored hydrogen bonds had a time of less than 10fs, which is too short for accurate sampling (Figure 9).

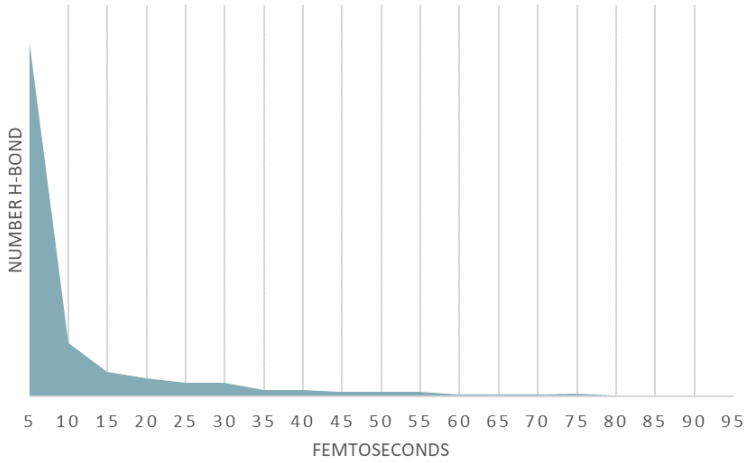

Figure 9 - Frequency of hydrogen bond duration

The focus has therefore shifted from the number of hydrogen bonds to their energies. A new parameter has been defined, called the water network variability (V_{H2O}), which is equal to the derivative of the hydrogen bond energies at simulation time:

$$V_{H_2O} = \frac{\sum_{res} \langle \left| \frac{\Delta E}{\Delta t} \right| \rangle}{i}$$

where ΔE is the change in energy of the hydrogen bonds over the time interval Δt, i is the number of water molecules.

By monitoring variability, we can estimate how water molecules interact with each other and how hydrogen bonds change. Molecules with low variability have a higher propensity to be replaced by a ligand. This may increase the number of partners available to establish more hydrogen bonds and increase their entropy.

15.2. Water mobility tracing via RMSF

To assess the degree of water mobility during the MD simulations, the positions assumed by each individual water molecule during the simulation were analysed by calculating the RMSF (Root Mean Square Fluctuations).

Water molecules with a low RMSF are molecules with little possibility of movement. The reason for this low mobility may be that interactions have been established between the molecules and the protein surface, whereas molecules with a high RMSF have not established strong interactions and their mobility is high. The entire trajectory of each water molecule that had a low RMSF was mapped.

Figure 10 shows the thyroid receptor beta 1 (PDB-id:1NAX) co-crystallised with the selective ligand KB-141. Of this complex, the molecular surface of the protein was coloured to show hydrophobicity and hydrophilicity (blue and yellow, respectively). It can be seen that the water molecules move towards the hydrophilic zones and, if they manage to establish enough interactions, remain there. The colours of the atoms and the connecting lines represent time (blue to red, with red corresponding to the last femtoseconds of the simulation).

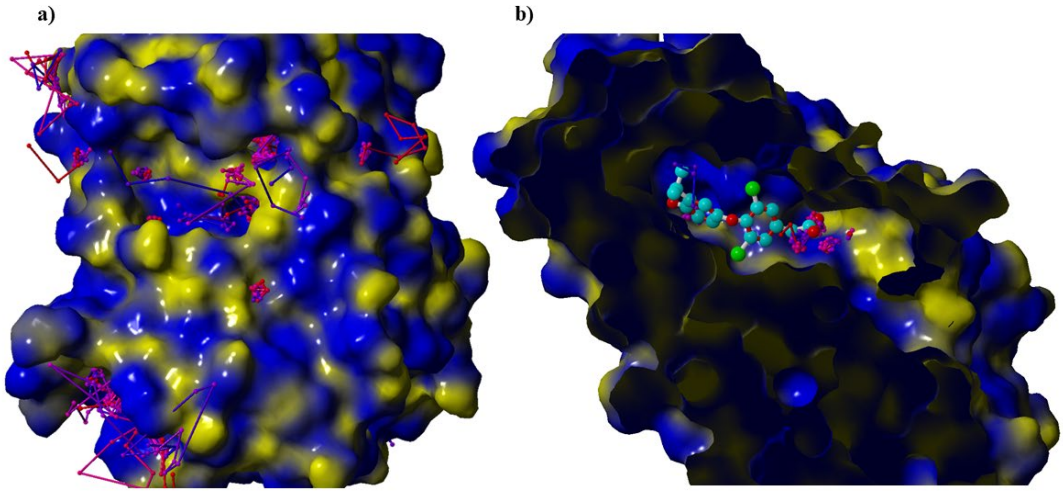

Figure 10 - a) Water mobility on the surface of the 1NAX protein. b) focus on water mobility of the receptor binding pocket. In both images, the surface with hydrophilic properties is indicated in yellow, the hydrophobic part in blue.

The choice of monitoring the movement of water molecules proved to be very interesting since it showed not only possible conformations of the binding pocket that can be exploited for new functionalisation of pre-existing ligands, but also the presence of allosteric sites.

15.3. WaterScope

Proteins achieve their biochemical functions by interacting with other biomolecules such as ligands, proteins, or nucleic acids. Identifying the binding site on a protein makes it possible to deduce the protein's function and provides information on binding pockets that are crucial for drug discovery. Over the years, several techniques have been used to predict the binding site, among them a software developed in the lab where I did my PhD research, YADA, that performs blind docking by checking for the presence of conserved portions in the protein structure [61].

Building on the studies of water molecules discussed above, a plug-in that could predict the position of the binding pocket was developed, called Water Scope. This is based on the idea that water molecules occupying the binding site have a greater propensity to be displaced by a ligand. Once displaced, the water molecules will have a greater number of partners with which to establish hydrogen bonds and increase their entropy. Consequently, the water molecules occupying the binding site have the least exchange of hydrogen bonds.

For the validation of this plugin, PDBs were selected from the BindingDB database [9], whose experimental structure and binding site are known. The structures were cleaned with the PDB-Clean plugin, and the ligand was removed.

MD was performed on each protein structure to solvate the entire system. The water molecules variability within 4Å of the protein was then calculated. The variability is calculated

as the derivative of the hydrogen bond energy for each water molecule during the simulation time. It is seen that the water molecules occupying the binding site have the lowest variability of all the molecules solvating the system (Figure 12).

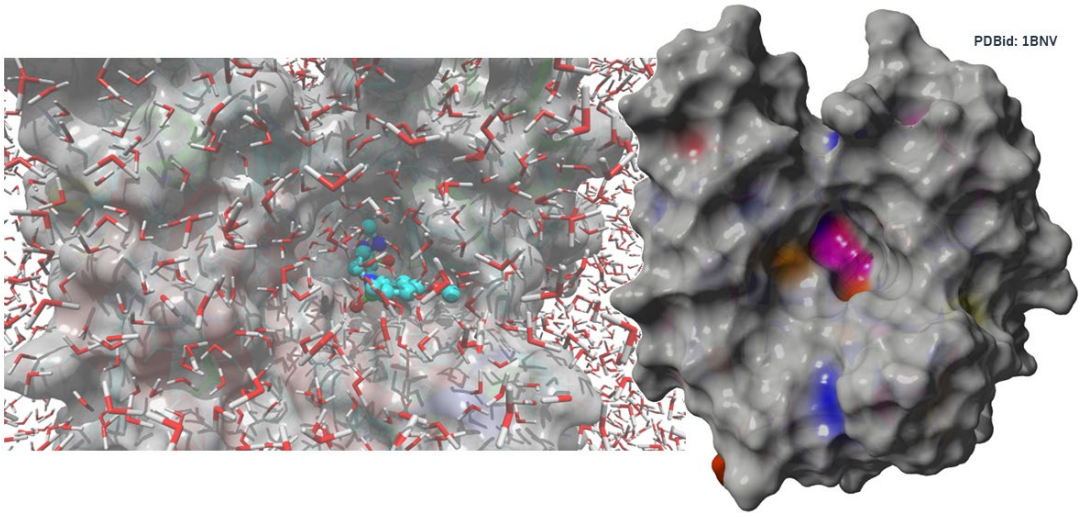

Figure 11 - On the left screenshot of the MD simulation where the System is fully hydrated. On the left, the result of the WaterScope plugin. The surface is coloured according to the variability of the water molecules surrounding it. In magenta the area where the water showed less variability, which highlights a possible active site. In this case, the magenta area perfectly identifies the active site.

15.4. Rethinking docking: the wrong model.

A typical docking software implements a sampling algorithm to generate possible binding poses and a scoring function to estimate their binding affinity. The first operation is known as pose generation, and the second is known as scoring. Over the years, many attempts have been made to associate a good scoring function with the generation of the correct pose.

Unfortunately, in most performance studies of docking algorithms, there is a core error. When evaluating how an algorithm predicts binding affinity, the energy associated with the best-predicted pose is considered. The best identified pose, however, does not always correspond to the experimental pose. For example, suppose the experimental binding energy of

a generic R-L complex is 12 kcal/mol and performing a docking experiment yields a prediction of 11.8 kcal/mol. In that case, the algorithm is thought to have a very good prediction. However, the predicted pose could not match the experimental one, and it is not possible to estimate the binding energy error from a wrong pose.

Based on this consideration, it was shown that choosing the wrong pose inevitably leads to an error in binding energy.

50 ligand-receptor complexes were selected for which the crystallographic structure of the complex and the experimental binding energy was available. Docking experiments were performed on each complex, and the relationship between the pose prediction and energy prediction errors was analysed (Figure 12).

In Figure 12, the abscissa is the RMSD between the experimental and predicted pose, and the ordinate is the binding energy error between the experimental and calculated pose in kcal/mol.

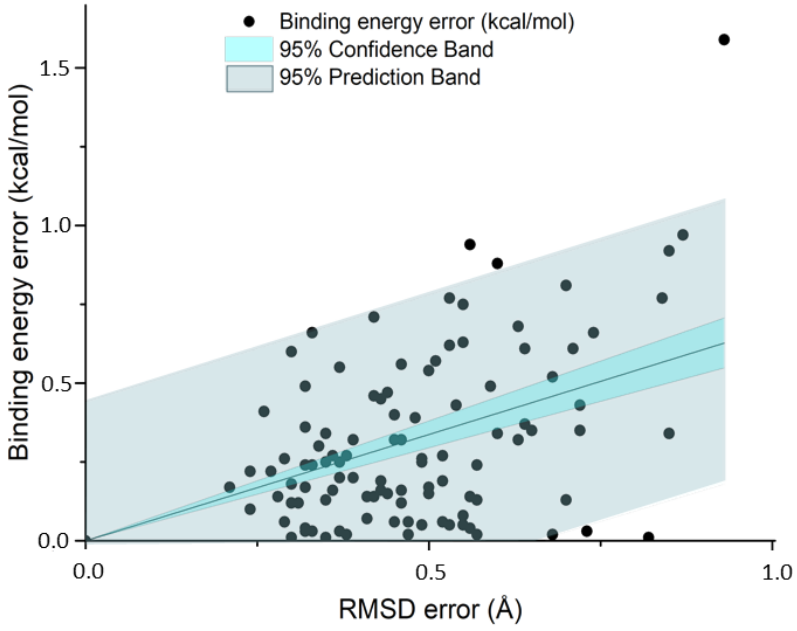

Figure 12 - Estimation of the error in the prediction of free binding energy in relation to the pose

The graph shows that an error of only 0.5Å in the prediction of the pose can lead to an error in the energy estimation of up to 1 kcal/mol, which could discriminate whether the analysed ligand is a good binder or not. As the RMSD increases, the error in binding prediction also increases.

For an RMSD with values greater than 0.5Å, we observe that there are very few values that fall within the confidence band, with values that do not even fall within the prediction band. these values suggest that the prediction model for the energy is inaccurate.

Therefore, it is clear that to have a good estimation of the binding energy of a complex, an accurate prediction of the ligand pose is necessary. This consideration was the starting point for this PhD project.

16. Results

16.1. Pose validation through RMSD calculation

Once the starting dataset was selected and cleaned with the PDBClean plugin, the pose prediction accuracy was evaluated with AutoDock and Vina. The RMSD between the best predicted pose and the experimental pose was predicted.

The results are shown in the following graph (Figure 13), in which the ROC curves and their respective AUCs are shown.

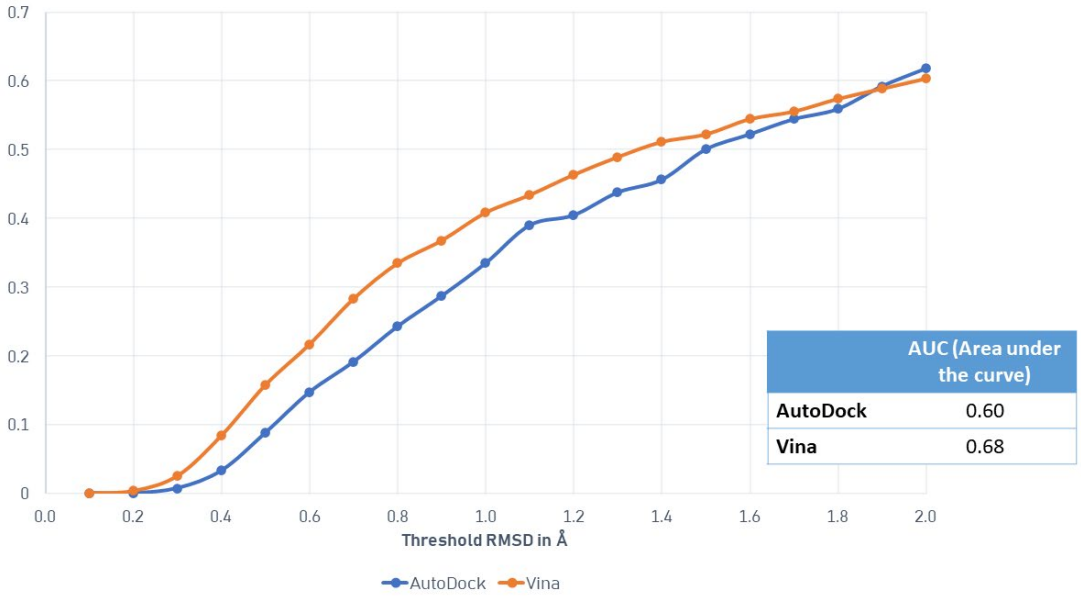

Figure 13 - ROC curves derived from the calculation of the RMSD between the experimental and predicted pose using the AutoDock and Vina algorithms

The AUC value for Vina is higher than Autodock, indicating a better performance of the algorithm in predicting the correct pose. The data showed that the RMSD is lower for the poses predicted using Vina.

16.2. Binding energy predictions

For the complexes that presented an RMSD lower than 2Å, the energy prediction was evaluated using the correlation (R^2) between the predicted and experimental binding energy. The results are shown in the scatter plot (Figure 14).

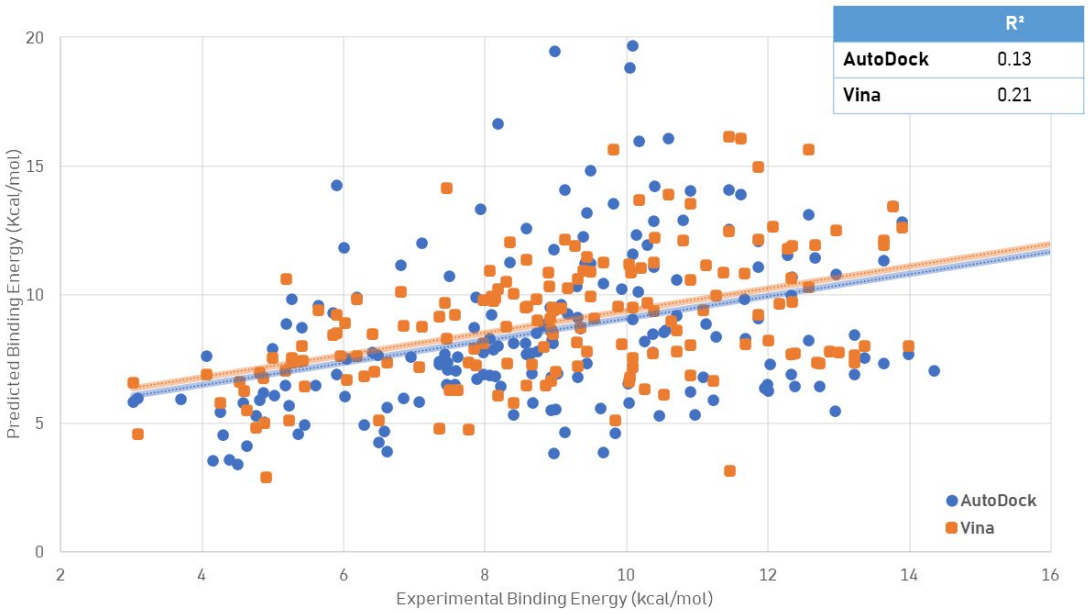

Figure 14 - Scatterplot showing the correlation between the experimental binding energy (in abscissa) and that predicted by the AutoDock and Vina docking algorithms (in ordinate). The value of R^2 is shown in the table at the top right.

The abscissa shows the experimental binding energy and the ordinate the energy predicted by the two algorithms. The correlation was evaluated using R^2. Vina showed a correlation of 0.21, higher than Autodock 0.13, confirming greater accuracy in predicting binding energy. The mean error was then calculated with its standard deviation shown in Table 2.

Table 2 - Mean error and standard deviation of binding energies predicted with AutoDock and Vina

	Mean Error (kcal/mol)	St. Dev
AutoDock	2.53	2.18
Vina	2.13	1.60

Vina has a smaller error than Autodock, but still high enough to be considered a good docking algorithm as the 2kcal/mol difference can discriminate between a good and bad binder.

Considering Vina's better performance in predicting both pose and binding energy, it was selected as the base algorithm for creating the new mathematical model.

16.3. Kinetic parameters of water

The choice of monitoring the movements of water molecules proved to be very interesting since it showed a new structural mapping of the surface and gave the possibility to formulate some considerations about the different degrees of mobility of water and how it could influence the calculation of the binding energy. Consequently, calculations were performed on the properties of water molecules solvating the binding pocket.

Molecular dynamics was performed by hydrating the entire system and the following parameters were calculated:

- Number of water molecules solvating the binding pocket
- The energy related to solvation water
- Water network variability (described in 10.4)

Energy and kinetic parameters of the water molecules were used as input data for the generation of the new mathematical model for calculating the binding energy, together with the Vina binding energies.

16.4. Generating a new model

Calculations on water molecules can be used to correct the binding energy estimate using genetic algorithms (GA). We start with a population consisting of a given number of possible solutions, and the genetic algorithm evolves these solutions. In the end, the algorithm selects the most suitable solutions and recombines them to obtain the best correlation. The software used is Materials Studio [60].

The selected target parameter is the experimental binding energy, and the input parameters used for model generation are:

- Value of ΔG_{bind} predicted by Vina
- Water molecule variability
- Average energy of water molecules
- Number of water molecules

The equation generated by the GA, called NewΔG for simplicity, is given below:

$$NewΔG = 0.082 * (Energy\ Water\ in\ Rec) + 0.023 * (Vina)^2 - 0.0002 * (Energy\ Water\ in\ Rec)^2 + 1.514$$

The new equation for predicting the binding energy was reapplied to the training set, and the correlation was re-evaluated (Figure 15).

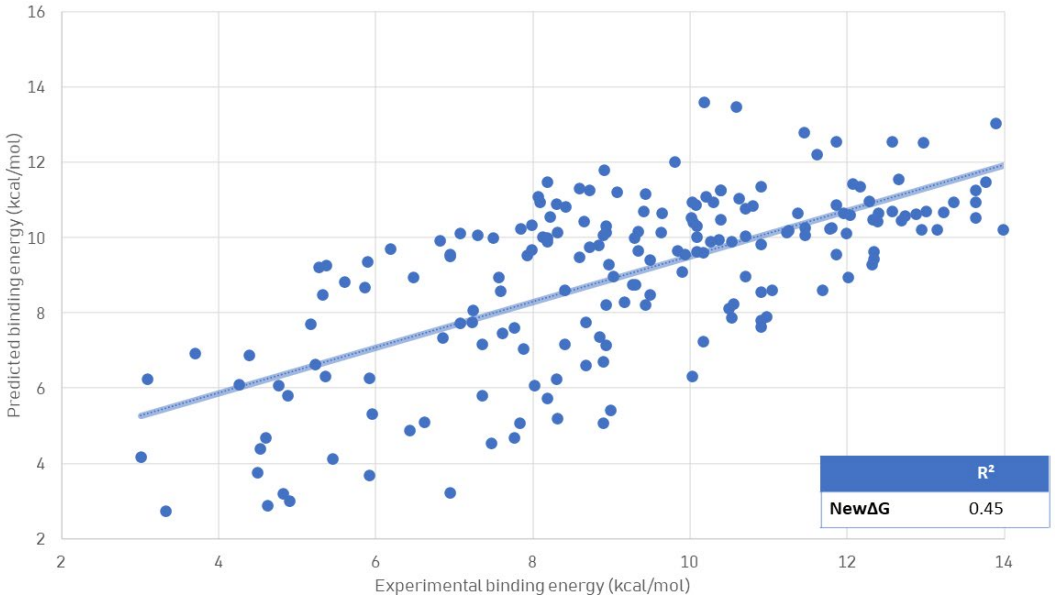

Figure 15 - Scatterplot showing the correlation between the experimental binding energy (in the abscissa) and that predicted using the new mathematical model (in the ordinate). The value of R2 is shown in the table at the bottom right.

The correlation between the predicted and experimental binding energy, calculated on the initial dataset, showed an R^2 of 0.45.

A plugin was then created, which interfaces with the Yasara software, which allows the binding energy to be calculated using the new equation. The code has been optimised to reduce the calculation time; in fact, a docking experiment on a complex takes about 30 seconds. Performance is based on a PC with Intel® Core™ i9-9900K CPU @ 3.60GHz and NVIDIA Quadro P4000 graphics card.

16.5. Model Validation

To validate the model, it was applied to a new dataset of 70 receptor-ligand complexes, and its performance was compared with AutoDock and Vina. The results are shown in Figure 16.

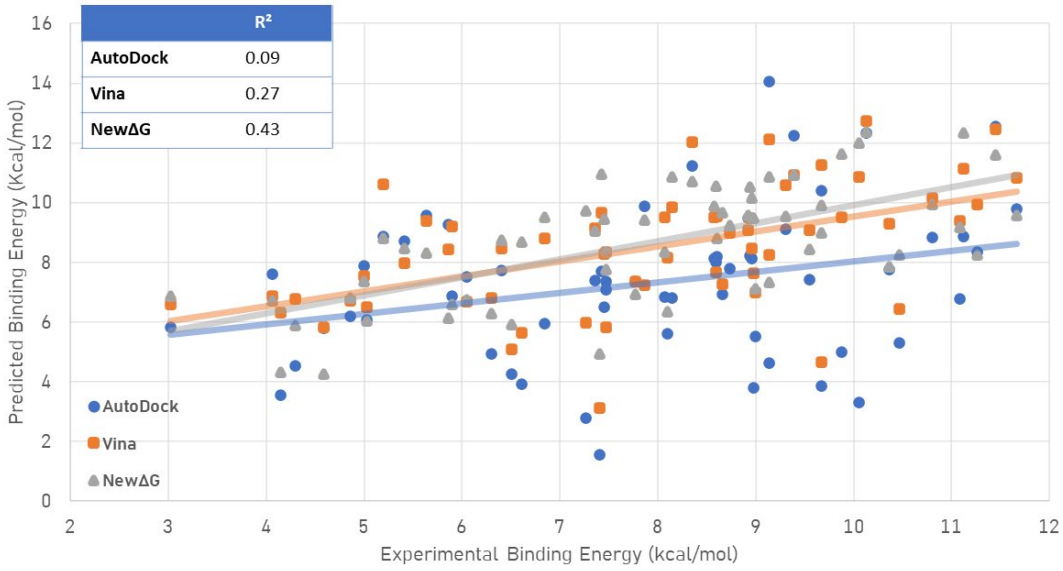

Figure 16 - Scatterplot showing the correlation between the experimental binding energy (in the abscissa) and those predicted with AutoDock, Vina and the new NewΔG (in the ordinate) calculated on the validation dataset. The corresponding R2 is shown on the top left.

The abscissa shows the experimental binding energy, the ordinate the binding energy predicted by AutoDock, Vina and the new model. The correlation was calculated, and it was

seen that the new model has an R^2 of 0.43, which is better than the other algorithms. The average error was also calculated and is shown in Table 3.

Table 3 - Mean error, the standard deviation of binding energies predicted and AUC

	Mean Error (kcal/mol)	St. Dev	AUC (Area under the curve)
AutoDock	2.23	1.70	0.62
Vina	1.68	1.23	0.70
NewΔG	1.45	0.86	0.75

The calculation error was reduced to 1.45 kcal/mol, comparable to the experimental error. These results are promising but not sufficient to determine the goodness of a predictive model. For this reason, the AUC (area under the curve) was calculated, indicating how well the model can discriminate false positives.

In general, the closer the AUC (Area Under the Curve) of the ROC is to 1, the higher the discriminating ability of the model. The AUC of the new model is 0.75, confirming the validity of the model.

PART I CONCLUSIONS

17. Conclusions

The primary goal of drug design is to describe the binding interactions between a drug and its target. Although pioneering studies in flexible docking and free energy computation are making significant progress towards improving the accuracy of docking and virtual screening regimes, these technologies remain complex, time-consuming, and still suffer from major errors and limitations. Changes in docking protocols are also driven by the evolution of artificial intelligence and machine learning algorithms for scoring and pose evaluation. With the increasing availability of experimental data, the field of molecular docking is witnessing the emergence of hybrid approaches trying to overcome all the limitations associated with it. Many current methods extend their strategies on machine learning principles, and it is in this field that my PhD project has entered.

The research work carried out as part of the three-year PhD in Drug Discovery and Development has led to developing a new docking protocol, implementing molecular dynamics and machine learning techniques to address the limitations described.

In the first part, the starting dataset with known experimental data was selected. The 3D structures were processed automatically using the PDB-Clean plugin. The initial dataset of 300 PDB was then divided into a training set and a validation set. The training set was in turn subdivided into other smaller datasets to avoid overfitting when using the genetic algorithms.

The input data for the genetic algorithms are derived from the docking experiments carried out on the training set and the energy and kinetic parameters derived from the molecular dynamics of the water molecules solvating the binding pocket.

The advantage of using genetic algorithms is that they evolve automatically, returning as output a mathematical model with the best correlation to the target parameter, in this case, the free energy of binding.

The obtained model was applied to the validation set to test its performance. The result is very promising since a correlation $R^2 = 0.43$ was obtained compared to $R^2 = 0.04$ for AutoDock and $R^2 = 0.20$ for Vina and a prediction error reduced to 1.4 kcal/mol, i.e., in the order of the experimental error. In addition, the area under the curve (AUC) was assessed to determine the goodness of fit of the model. The new model has an AUC of 0.75.

Based on the very promising results obtained, we will try to expand the learning dataset in the future to obtain an increasingly robust model.

PART II:

Study on the role of 2-hydroxyoleic acid and Sphingomyelin Synthase

INTRODUCTION

1. Glioblastoma

About 70% of all tumours affecting the central nervous system (CNS) are gliomas [62, 63], a group of brain tumours originating from cells in the glial line. Glial cells not only provide mechanical support to neuronal cells in terms of nutrients, oxygen, and disposal of waste products, but are also involved in all the mechanisms of signal transduction and communication with the neuronal cells themselves.

Gliomas are classified histologically and immunohistochemically into astrocytomas, oligodendrogliomas and tumours with astrocytic and oligodendrocytic morphological features, called oligoastrocytomas. These tumours are then subdivided from grade I to grade IV according to their degree of malignancy and genetic alterations (Figure 17), accompanied by cell size, degree of cellular pleiomorphism, mitotic activity, degree of proliferation and necrosis of pericytes and microvascular endothelium [64].

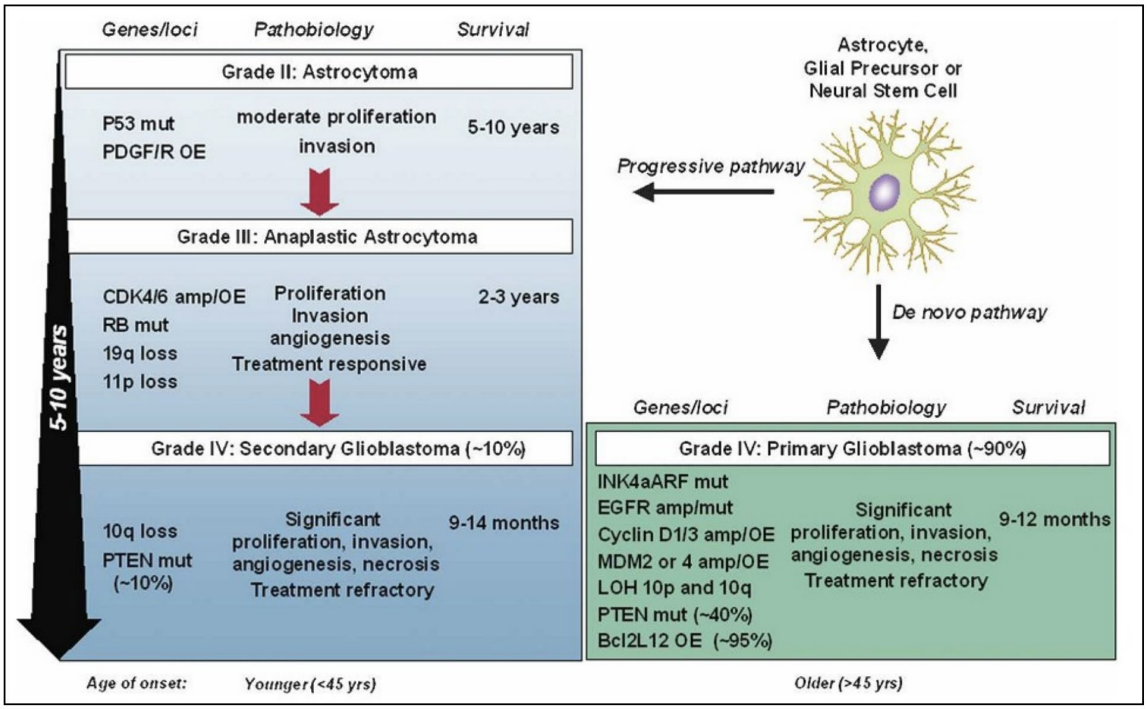

Figure 17 - Representative diagram of the various grades of glioma. Adapted from Furnari et al 2007, Ref: [64].

Grade I tumours are biologically benign and can be cured if removed in time by surgery. Grade II tumours are of low malignancy and have a potentially long clinical course; however, they are characterised by an early and diffuse infiltration event into the surrounding brain tissue, making them untreatable by surgery. Grade III tumours show more anaplasia and proliferation than grade II tumours and lead the patient to death more rapidly. Grade IV tumours are characterised by the highest degree of malignancy and involve vascular proliferation and necrosis events. Grade IV tumour is also called glioblastoma multiforme (GBM) or, more simply, glioblastoma [65].

GBM is the most common and aggressive glial tumour, as it tends towards invasiveness and cell proliferation. The incidence is 2-3 new cases per year per 100,000 inhabitants. The average survival is 14 months, even after aggressive surgical removal combined with radiotherapy and chemotherapy. Rare, but present, are cases of survival.

Like all brain tumours, except in rare cases, glioblastoma does not expand beyond the CNS structures. GBM can be divided into primary and secondary types. Primary GBM mostly affects elderly patients, whereas secondary GBM is rarer and occurs under 45 years of age. Although morphologically and clinically indistinguishable, primary and secondary GBM are characterised by marked differences in genomic, RNA and protein profiling, and response to chemo- and radiotherapy that reflect their different clinical histories [64]. Among all known types of solid tumours, glioblastoma multiforme is the one with the highest degree of angiogenesis. Understanding the molecular mechanisms that drive the angiogenic process in GBM is of fundamental importance to study therapies that block its progression.

In general, specific anticancer therapy for GBM involves (not necessarily in this order):

- surgery
- radiotherapy
- chemotherapy with temozolomide.

Sometimes GMB is initially inoperable, and radiotherapy and/or chemotherapy is used in the first instance to reduce the tumour mass and allow the neurosurgeon to intervene later. If the type and location of the tumour allow it, a complete macroscopic resection of the tumour lesion is always performed by surgery. Although this multimodal therapy has improved patient outcomes in recent years, improvements to current approaches and/or alternative therapies are urgently needed [66].

Some anticancer drugs appear to act by regulating signal transduction and altering the lipid structure of the plasma membrane [67]. Anthracyclines such as doxorubicin, epirubicin and idarubicin are possible alternatives to temozolomide-based chemotherapy. These cytotoxic compounds, derived from Streptomyces bacteria, are effective against various cancers due to their ability to induce apoptosis in cancer cells [68, 69].

Anthracyclines are characterised by a rigid planar aromatic ring linked to an amino sugar. The quinone groups allow the exchange of electrons in the conversion of quinone to semiquinone radical [70, 71]. This radical is converted back to quinone under aerobic conditions, resulting in the formation of superoxide anion and hydrogen peroxide. This leads to the excessive formation of free radicals, resulting in peroxidation of lipids within cell membranes, DNA damage and ultimately cell death [72].

Due to this mechanism, anthracyclines are considered potent non-selective anti-cancer drugs, and are used in the treatment of a wide range of cancers [73-75].

Anthracyclines are are prescribed to more than 30% of breast cancer patients and more than 50% of all patients with childhood cancer [76]. In particular, the efficacy of doxorubicin against solid tumours is well established, and it is listed as an essential medicine by the World Health Organisation (WHO) [77].

Although anthracyclines have shown a powerful effect in inhibiting cell growth in many types of tumours, including CNS neoplasms [78, 79], most patients fail to achieve adequate

disease control due to limited drug penetration into the CNS, as Von Holst and co-workers have shown in patients with malignant gliomas [80].

Anthracycline-based chemotherapy regimens are not currently used to treat solid intracranial tumours like GBM, as they do not pass the blood-brain barrier [81]. To overcome this severe limit, a compound was designed that could reproduce the antitumour effect of anthracyclines through interactions with the plasma membrane and the resulting changes in cell signalling [82], without non-specific interactions with other cellular targets: 2-hydroxyoleic acid (2OHOA).

2. LP651 as a new anti-cancer drug

The anticancer compound 2-hydroxyoleic acid (LP561) acts against cancer by inducing cell cycle arrest [83], followed by apoptosis in human leukaemia cells [84] or differentiation and autophagy in the case of human glioma cells [85]. Given the potency of 2OHOA against cancer, it has been shown to be a safe and non-toxic compound with IC50 values in non-tumour cells 30 to 150 times higher than in cancer cells [84]. This fatty acid's high efficacy and low toxicity produce a wide therapeutic window that can only be the consequence of a highly specific mechanism of action, the molecular basis of which has not yet been fully elucidated.

2OHOA binds and modifies the biophysical properties of the lipid bilayer, the first target encountered by this synthetic lipid [86]. We know that this molecule induces changes in the localisation and activity of membrane proteins involved in the proliferation, differentiation and survival of tumour cells, such as the Fas receptor [84], PKC [87], as well as cyclins, cyclin-dependent kinases (CDKs), caspases, E2F-1 and dihydrofolate reductase (DHFR) [83]. Interestingly, a similar mechanism is described for edelfosine, a synthetic ether lipid with a high apoptotic activity that also induces membrane raft reorganisation, FasR capping and cell apoptosis [88].

Membrane lipid therapy is a new therapeutic approach in which drugs are designed to target the membrane of diseased cells, modulating its composition and structure, and thereby modifying the protein activity that interacts with the membrane. 2OHOA is specifically designed to regulate the lipid composition and structure of the cell membrane (membrane lipid therapy).

Additionally, 2OHOA produces significant changes in the composition of membrane lipids by increasing the synthesis of sphingomyelin (SM) with the activation of sphingomyelin synthase (SMS).

Since 2OHOA has relevant antitumour activity against glioma, and since its mechanism of action involves SMS activation and altered SMS1 expression is associated with survival of glioma patients, the role of these enzymes in glioma carcinogenesis, prognosis, and response to 2OHOA has long been investigated. It has been seen that 2OHOA regulates SMS activity in tumour cell membranes but not in those of non-tumour cells, and concomitant changes in SM and other lipid levels. Exposing cancer cells to 2OHOA promotes a robust increase in SM through the rapid and specific activation of SMS enzymes. The immediate activation of SMS by 2OHOA and subsequent accumulation of SM may at least partly explain the ability of this compound to trigger cell cycle arrest, cell differentiation and autophagy or apoptosis in cancer cells. Importantly, 2OHOA has differential and specific effects against cancer cells.

3. The sphingomyelin synthase family

In recent years, more and more evidence has accumulated showing a possible relationship between plasma membrane sphingomyelin levels and tumorigenesis. SM is synthesised by sphingomyelin synthase (SMS), which catalyses the transfer of a phosphocholine head group

from phosphatidylcholine (PC) to a ceramide (Cer), producing a diacylglycerol (DAG) and SM [89] according to the reaction:

$$PC + Cer \leftrightharpoons SM + DAG$$

Sphingomyelin is a key component of the plasma membrane that interacts with cholesterol and glycerophospholipids, thus participating in the formation and maintenance of lipid microdomains. Lipid rafts are important signalling platforms whose structure is sensitive to the composition of membrane lipids [90], as are the proteins that interact with these and other microdomains [91]. Thus, changes in SM content affect the signalling pathways associated with the lipid raft. The SM is located at the junction between the two main groups of membrane lipids (glycerophospholipids and sphingolipids) and between two key signalling molecules in cell cycle regulation (ceramide and 1,2-DAG) [92]. This process can occur in the Golgi apparatus mediated by SMS1 or in the plasma membrane by SMS2 [93]. The two isoforms share 57% sequence identity and are conserved in mammals [93, 94]. SMS1 contains a sterile alpha motif (SAM), involved in protein-protein interaction, which is not present in SMS2. SMS are related to the lipid phosphatase family (LPP), with six transmembrane regions with N- and C-terminals exposed in the cytosol [95]. Although SMS requires PC as a head group donor to form SM, overexpression or knockdown of SMS mainly affects sphingolipid levels (SM and ceramide) without noticeable changes in PC levels [96, 97].

Since SMS can reverse SM production by forming SM and DAG again, they are also considered regulators of proapoptotic ceramides and DAG as a second messenger [98]. In this regard, it has been seen that SM levels are reduced in a variety of tumour cells compared to non-tumour cells and restoring normal SM levels by activating SMSs inhibits tumour cell proliferation and/or induces cell death [98]. As SMS and sphingomyelinases are linked to multiple diseases, some authors predict [99] that more drugs targeting the SM cycle will be developed in the future.

4. The role of sphingolipid hydroxylation

Sphingolipids are key molecules in regulating the cell cycle, apoptosis, angiogenesis, stress, and inflammatory responses. A further feature in all classes of sphingolipids that is also shared with other phospholipids is that they can be hydroxylated [100]. Hydroxylation of sphingolipids, both in the acyl chain and in the sphingolipid backbone, may also influence membrane lipid packing and G-protein regulation [101]. Hydroxylation patterns greatly influence the biophysical properties of sphingolipids, as illustrated, for example, by the significant difference in the disordered gel-liquid phase transition temperature (Ld) when comparing similar sphingolipids with different hydroxylation patterns [102, 103]. More important, perhaps, is the influence of hydroxylation in the interaction between sphingolipids and the surrounding membrane containing other lipid components. Recently, important studies have been reported on membrane interaction with sphingolipid compounds containing -OH [104]. Sphingolipids containing 2-hydroxylated fatty acids (2OHFA) are present in most organisms [102] and are important components of a subset of mammalian sphingolipids. The enzyme FA2H (fatty acid 2-hydroxylase) is a hydroxylase that introduces a hydroxyl group into the 2-position of fatty acids [105]. 2-hydroxy fatty acids are found almost exclusively as N-acyl chains in the ceramide fraction of various sphingolipids [106]. FA2H is stereospecific in producing (R)-2-hydroxylated fatty acids [107]. Hydroxylation at position 2 occurs during de novo synthesis of ceramide and is catalysed by FA2H [105]. In mammals, the six isoforms of CerS (Ceramide synthases) can use 2-hydroxy-acyl-CoA as substrates to synthesise 2OHFA-dihydroceramide [108]. In addition, it has been shown that galactosylceramide synthase has a strong preference for 2OHFA ceramide over non-hydroxylated ceramide [109]. The influence of hydroxylation can be further studied by comparing how hydroxylated, and non-hydroxylated lipids interact with related enzymes.

5. CLINGLIO

During the last year of this PhD project, I had the opportunity to participate in the European project CLINGLIO. It involved 17 partners across Europe (https://clinglio.eu/), including the University of Salerno and the University of the Balearic Islands, where I carried out part of my research under the supervision of Prof. Pablo Vicente Escribá.

2OHOA has shown positive results in the treatment of glioma in several studies, and it has been approved as an orphan drug for the treatment of glioma by the European Medicines Agency (EMA).

The European CLINGLIO project was involved in advancing clinical trials of 2OHOA for the treatment of glioblastoma and marketing in Europe. The CLINGLIO project included a phase IIB demonstration clinical trial to evaluate the efficacy of a novel therapy based on 2OHOA and SoC (standard of care) in newly diagnosed subjects with primary glioblastoma. In addition, planned studies with patient samples and glioma cells aimed to further characterise biomarkers for predictive (diagnosis and prognosis with threshold biomarkers and omics signatures), pharmacodynamic/pharmacogenomic and stratification (patient treatment allocation) purposes.

A randomised, double-blind, placebo-controlled, 2-arm parallel study (1:1 ratio) was planned to evaluate the efficacy of 2-hydroxyoleic acid (2-OHOA) compared to placebo in patients with newly diagnosed, wild-type IDH glioblastoma. In all arms, patients will receive SoC and will be randomised to receive placebo or 2OHOA dose.

2OHOA had already been shown to be safe in patients with glioma and other advanced solid tumours during Phase I/IIa clinical trials (ClinicalTrials.gov ID #NCT01792310). In a xenograft model of human GBM, 2OHOA provided a greater anti-tumour effect than temozolomide

(TMZ), which is the current standard first-line chemotherapy against GBM and increased patient survival by approximately 2.5 months.

AIM OF THE STUDY

6. Aim of the study

2OHOA is a first-in-class anti-cancer lipid due to its mechanism of inhibition of MAPK (mitogen-activated protein kinase) and related oncogenic pathways [85]. However, the molecular mechanism has yet to be fully elucidated. It has been experimentally demonstrated that 2OHOA activates the enzymes SMS, inducing a rapid increase in SM levels in the membrane.

In the second part of my PhD project, the plausible mechanism of action of 2OHOA on both SMS isoforms and the role of hydroxylated lipids in this scenario was investigated. Several computational techniques were applied to these systems, of which little experimental knowledge is available. A possible mechanism of action of 2OHOA with the SMS isoforms and a related energy profile has been outlined by estimating the free energy of binding in the intermediate stages. The existence of middle stages allows clarifying the role of hydroxylation on the carbon at position 2 in PC, Cer, and SM chains.

As the three-dimensional structures of this protein are not available in the literature, we started by predicting the 3D structures of SMS1 and SMS2 using various computational techniques, and identifying the possible binding site, probably located in the transmembrane region. Docking and molecular dynamics experiments were performed, which allowed us to identify a key role of a tyrosine present in the binding site (Tyr223 for SMS1 and Tyr167 for SMS2), which allows a possible nucleophilic attack to transform PC into SM. All energy profiles were delineated using docking and metadynamics experiments.

The obtained results suggest a role for another enzyme, ceramide synthase, in incorporating 2OHOA (particularly in the R form) into ceramide, which has been shown to be energetically favoured for interaction with the enzyme. These preliminary results pave the way for a better

understanding of the role of 2OHOA and, more generally, of hydroxylated sphingolipids in the mechanisms controlling autoimmunity in healthy individuals.

MATERIALS AND METHODS

7. Structure Prediction and Validation

The three-dimensional structures of the SMS isoforms are not known. The amino acid sequences are available and downloaded from the Uniprot database [110] (for human SMS1 code Q86VZ5, for human SMS2 code Q8NHU3).

SMS1 consists of 413 residues and is organised into two cytosolic fragments (N- and C-terminal) and the transmembrane portion. SMS2 consists of 365 residues, and the main difference with SMS1 is the lack of the SAM domain. Sequence alignment of the two isoforms was performed with the multiple alignment program Clustal Omega [111]. The tertiary structure was predicted de novo using the Folden modelling suite [112]. The algorithm uses a dedicated neural network to predict the inter-residual distance and orientation distributions of the input sequence. The models of the two SMS isoforms with the best TM-score have been validated by PROCHEK v.3.5 web server [113], a widely used web service for validating three-dimensional structures.

8. Binding Site Definition

Once the structure has been predicted, it is necessary to identify the protein's binding site. Two different approaches have been used: conservation strings and WaterScope. In the literature, it is known that conserved sites on the surface of the protein play a crucial role in the activity of the enzyme. Therefore, the conservation string was obtained from the Consurf database [114], a server to identify structurally important residues in protein sequences. The conservation string varies from 9 for highly conserved residues to 1 for non-conserved amino acids, as described in ref. [61].

The WaterScope plugin [112], described in section 10.4, is set up with a 5 Å cuboid simulation cell around the receptor. The coordinates of each receptor atom were fixed, and the system was neutralised with NaCl at a concentration of 0.9%. The charges were assigned at pH 7.0 by applying the force-field AMBER15IPQ [115]. The system was neutralised respecting the density of water at 0.997 g/mL [116]. The Berendsen thermostat was applied at 298 K with a timestep of 1.25 fs. After the neutralisation phase of the system, a 50 ns molecular dynamics simulation was performed using the SolventProbe barostat available in the Yasara software. Changes in the hydrogen bonding network of water molecules around the entire surface of the protein were monitored.

9. Molecular Docking

To predict the geometry of the complex of the two SMS isoforms with the natural and hydroxylated substrates, docking experiments were performed using VINA as an algorithm [42] and Yada software [61]. The use of these two software packages allows to reach a consensus on both the geometry of the laying and the energy calculation. AMBER14 [117] force-field was used for both software. With VINA, the ligands were independently docked 250 times with 5 receptor ensembles. The simulation cell was defined around the key residue Tyrosine 223 for SMS1 and Tyr167 for SMS2 (the reason for this choice is explained by the results obtained from molecular dynamics). The results were clustered with an RMSD of 5.0 Å. With Yada, we used the same blind docking procedure as described in these works [118, 119]. We chose 250 runs per hotspot (the centre of gravity of the conserved residues) in a box 20 Å larger than the receptor. Natural substrates of the enzyme (PC, SM, Cer, and DAG) and their hydroxylated forms were considered as ligands for a total of 12 ligands (Figure 18).

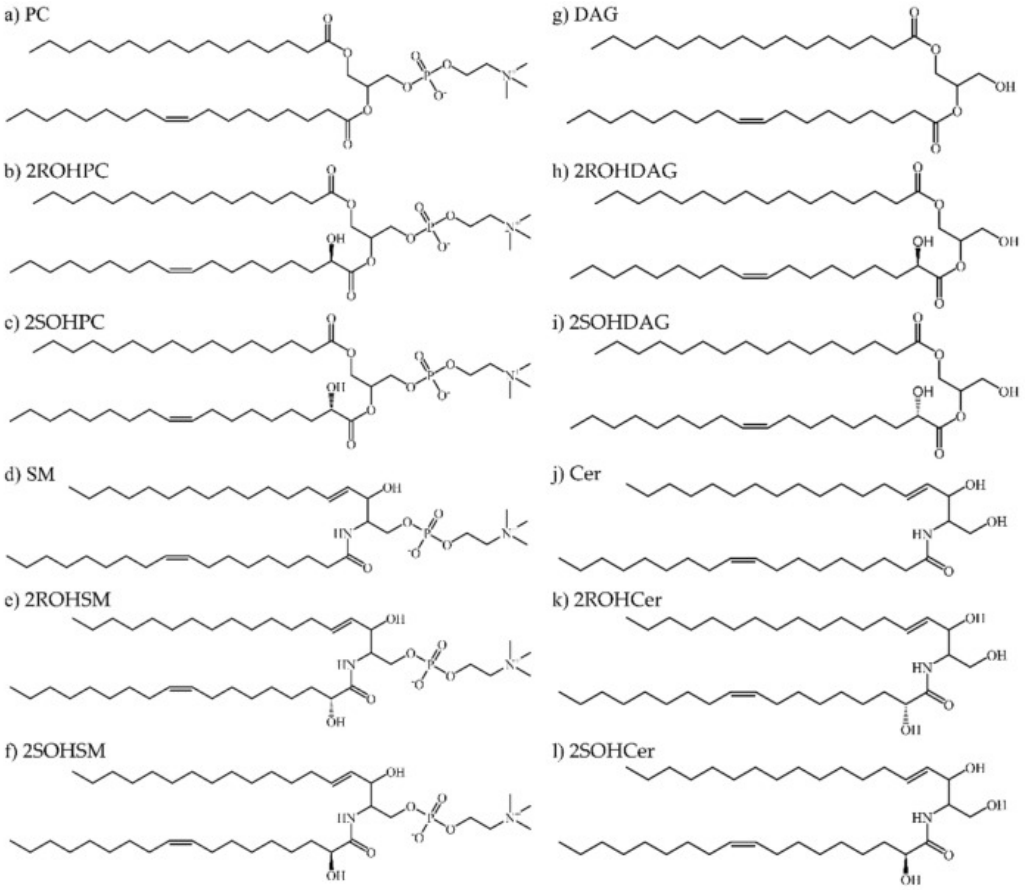

Figure 18 - Chemical structures of the ligand set. (a) 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (PC), (b) hydroxylated PC with 2ROHOA (2ROHPC), (c) hydroxylated PC with 2SOHOA (2SOHPC), (d) N-oleoyl-d-erythro-sphingosylphosphorylcholine (SM), (e) hydroxylated SM with 2ROHOA (2ROHSM), (f) hydroxylated SM with 2SOHOA (2SOHSM). (g) (S)-1-hydroxy-3-(palmitoyloxy)propan-2-yl oleate (DAG), (h) hydroxylated DAG with 2ROHOA (2ROHDAG), (i) hydroxylated DAG with 2SOHOA (2SOHDAG), (j) ceramide N-((2S,3R,E)-1,3-dihydroxyoctadec-4-en-2-yl)oleamide (Cer), (k) hydroxylated Cer with 2ROHOA (2ROHCer), (l) hydroxylated Cer with 2SOHOA (2SOHCer).

10. Molecular Dynamics Simulations

The stability of protein-ligand complexes in membranes was analysed by using molecular dynamics simulations. Docked poses of the protein-ligand complexes were used as input structures, and each complex was pre-processed with the Desmond [120]. First, protein-binding complexes were pre-processed using the protein preparation wizard in the Maestro 2021-1 suite obtained through academic licensing. Missing hydrogens were added, the bonding orders were

assigned, and the protein was minimised using the OPLS3e force field [121]. The systems were centred in an orthorhombic box with the edges 10 Å from the protein. The complex was immersed in a PC membrane, using it as a model membrane. The orientation of SMS1 and SMS2 in the membranes was predicted using the OPM web server [122]. The two proteins had a similar tilt angle (23 ± 1 ◦ for SMS1 and 25 ± 1 ◦ for SMS2). The system was solvated (Tip3P water model) and neutralised with Na+ and Cl- ions. The salt concentration was set to 0.15 M to simulate physiological conditions. MD was conducted under periodic conditions in the NPT ensemble using the OPLS3 force field. Temperature and pressure were maintained at 300 K and 1013 bar, respectively, using Langevin temperature coupling and isotropic scaling. The simulation time is 30ns. Subsequently, the trajectories were analysed to monitor the interactions of the ligand atom with the protein residues and the protein interactions with the ligand.

11. Metadynamics Simulations

The phosphorylated form of SMS (SMS1-P and SMS2-P) embedded in a PC membrane was used for the metadynamics simulations. GPU-accelerated Desmond software was used on an NVIDIA GeForce GTX 980 graphics card. , using the Langevin thermostat and barostat. A combination of two collective variables (CVs) describing the movement of ceramide (and hydroxylated analogues) into the phosphorylated protein binding site was defined. For distance CVs, the width of the Gaussian was set to 0.05 Å. The starting height of the Gaussian potential was set to 0.03 kcal/mol, and Gaussians were deposited every 0.09 ps. Simulations were performed at 300 K and 1.013 bar pressure. The RESPA integrator was used with a timestep of 2.0 fs. For coulombic interactions, the cut-off was defined at 9 Å. No position constraints were specified for any atom. Trajectory frames were recorded at an interval of 20 ps for a simulation

time of 30 ns. The simulations were displayed in the Maestro suite [123]. Trajectory analysis was performed using Maestro's Simulation Event Analysis and Visual Molecular Dynamics.

RESULTS AND DISCUSSION

12. Discussion and Results

12.1. Prediction and validation of the 3D structure of the two isoforms

There are two main isoforms of SMS, which share high sequence homology, except for a cytosolic SAM (Steril Alpha Motif) domain that is absent from the SMS2 isoform. Isoform 1 consists of 413 residues, and isoform 2 of 365 amino acids. It was seen that the transmembrane portion (residues 131-353 for SMS1 and 74-294 for SMS2) has a similarity of 74.55%.

In [94], the structure of SMS1 had already been reconstructed using homology modelling. However, in recent years, there have been considerable improvements in algorithms for predicting the 3D structure of a protein, many based on the use of dedicated neural networks. For this reason, it was decided to re-predict the 3D structure using the Folden software in the SMP modelling suite [112]. The 3D models were submitted to the PROCHECK server, where Ramachandran plot statistics were generated (Figure 19).

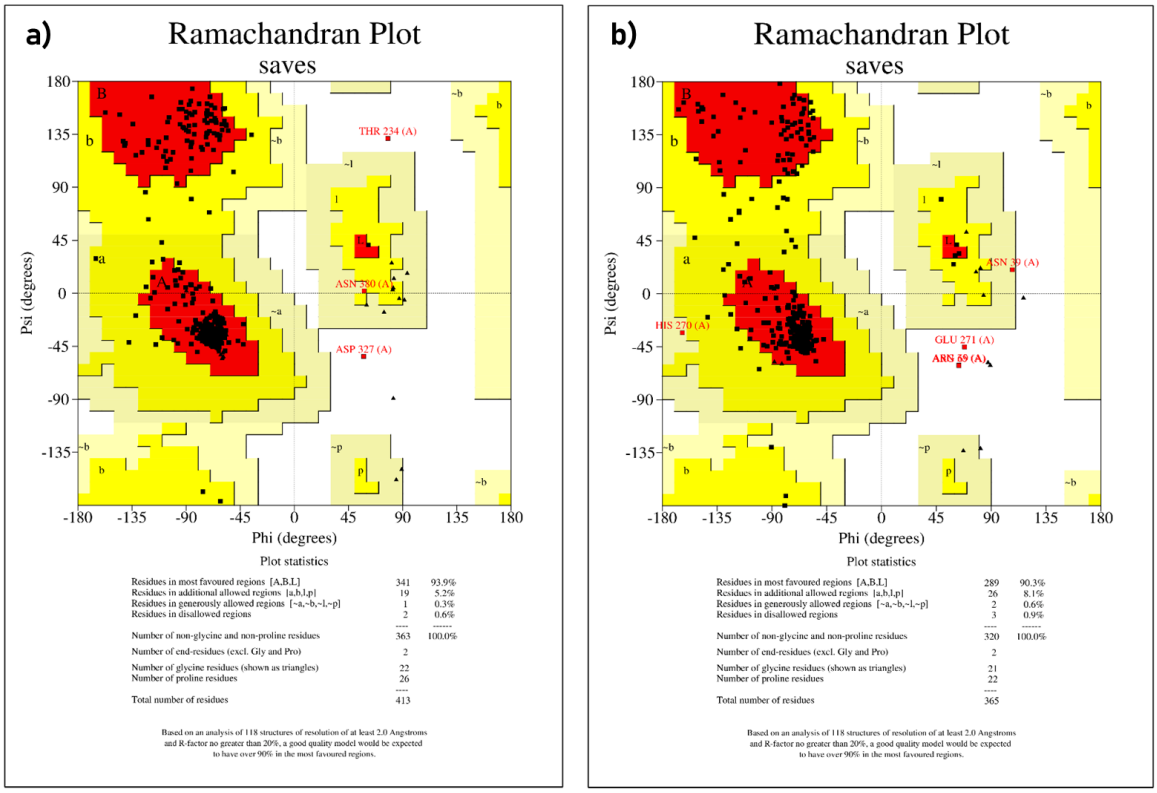

Figure 19 - Ramachandran plot for SMS1 (a) and SMS2 (b) showing the presence of amino acid residues in favoured, allowed, and outlier regions.

For SMS1, the output showed 93.9% residues in the most favoured region, 5.2% residues in the permitted supplementary region, 0.3% residues in the generously permitted regions and only 0.6% residues in the non-permitted regions. For SMS2, the output showed 90.3% residues were present in the most favoured region, 8.1% residues in the permitted additional region, 0.6% residues in generously permitted regions and 0.9% residues in non-permitted regions.

The two SMS isoforms showed a sequence homology percentage of 62.07% over the entire structure, but the cytosolic portions were less conserved than the transmembrane portion. Therefore, we only compared the transmembrane portion. In accordance with the prediction of Phobius [124], for SMS1 we selected 222 residues from residue E131 to residue Q353, and residue E75 to residue E297 was selected for SMS2. In detail, the two transmembrane portions shared 91.9% fully conserved residues and residues with strongly similar properties. Using

docking and molecular dynamics studies, we sought to define whether these slight differences could influence the type of interactions at the binding site.

12.2. Binding site identification

In the literature [125, 126], it has been hypothesised that the most conserved portions in the two proteins play a crucial role in the enzyme's activity. Consurf was used [114] to estimate the evolutionary conservation of amino acid positions in a protein-based phylogenetic relationship between homologous sequences [127]. The magenta-coloured regions in Figure 20 are the most conserved regions between the two proteins and indicate the highest probability that this is a binding site.

An active site can be more accurately predicted by monitoring the movement of water molecules during molecular dynamics. For a ligand to interact with a receptor, the water molecules present at the binding site must be moved. This concept is the basis of the WaterScope tool, described in detail in section 15.3.

As shown in Figure 20, the core of the transmembrane portion corresponds to the position of the water molecules with low mobility. Stationary molecules in the transmembrane locate the catalytic site in this portion. The portion highlighted with a blue surface corresponds to the highly conserved transmembrane portion shown in the figure above.

The SAM domain is also a catalytic domain, but SMS mutants with deletion of the SAM domain from SMS1 showed no significant impact on SMS catalytic activity [128]. Therefore, later studies focused on the transmembrane portion. The natural substrates of SMS, PC and SM are molecules that are not very soluble in water and can only access the membrane-immersed portions of the enzyme. The membrane components can easily reach the binding site in the transmembrane portion by moving freely across the membrane.

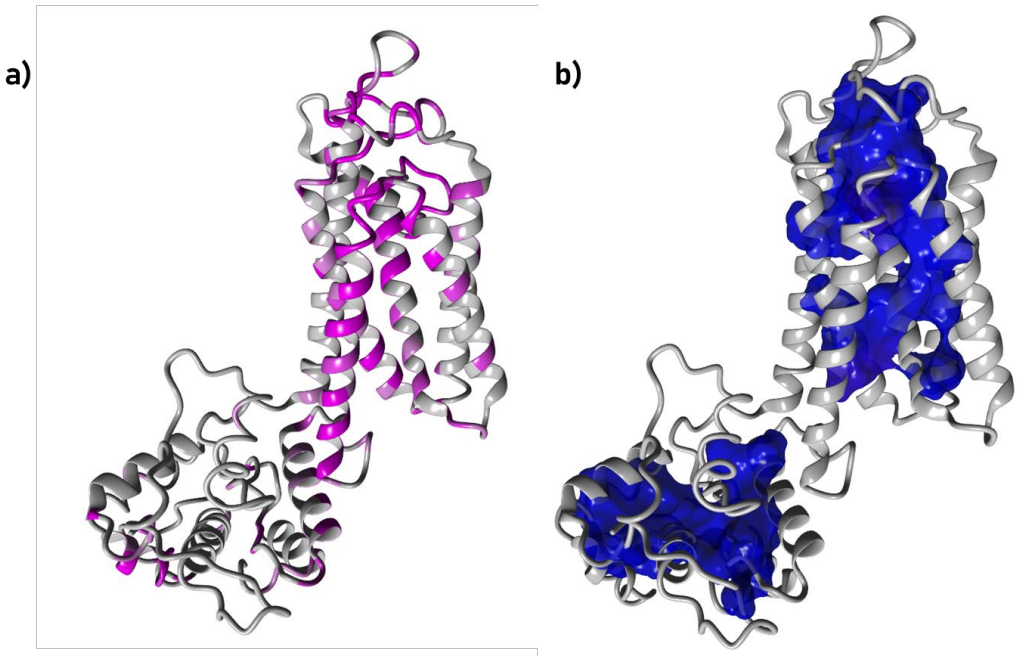

Figure 20 - Identification of potential binding sites. Highlighted in magenta (a) are the conserved sites for SMS1 and in blue (b) is the surface of SMS1 highlighted by waterscope

12.3. The key role of tyrosines

Docking experiments were performed using the AutoDock VINA algorithm [42] and YADA software [61] on both isoforms of SMSs with their natural PC substrate, placing the simulation cell around the transmembrane region to evaluate the binding energies and the interactions between the ligands and receptors. The best fit was used as an input structure for molecular dynamics (MD) simulations for each isoform. The complex was embedded in a POPC model membrane.

Molecular dynamics trajectory analysis for SMS1 revealed contact for 99% of the simulation time between Tyr223 and the phosphate group of the PC substrate. This interaction suggested a possible nucleophilic attack of the hydroxyl group of Tyrosine 223 to the phosphorus atom of the PC molecule. In this way, after the formation of the O-P bond between

the phosphocholine head and the tyrosine of the enzyme, the DAG molecule can then move away from the active site by crossing the membrane (Figure 21).

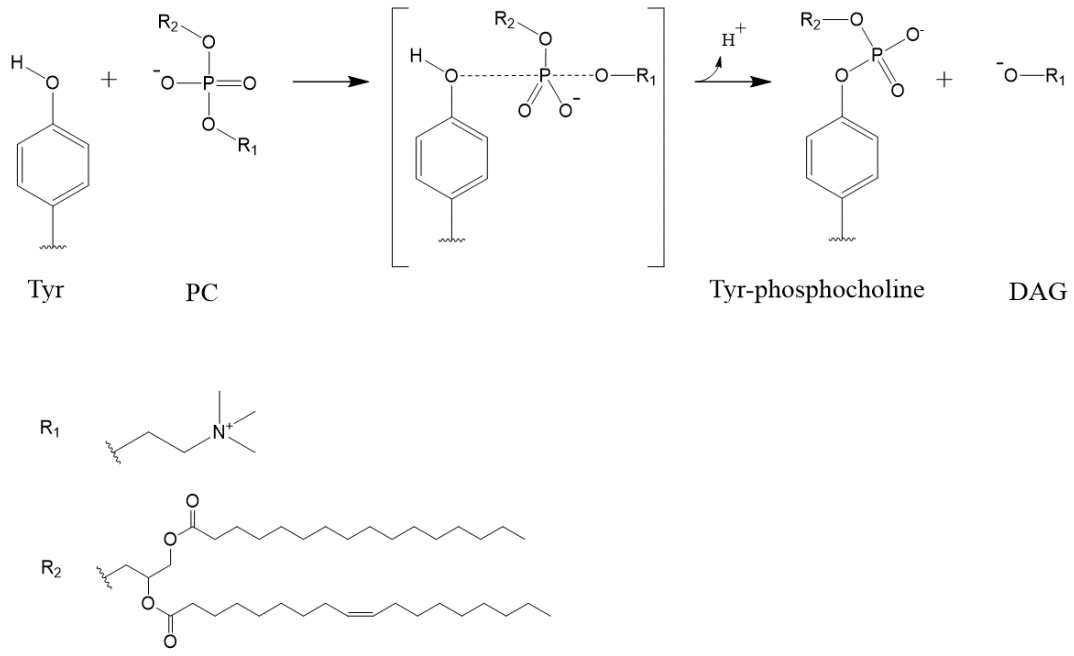

Figure 21 - Hypothesis of the first step of the SMS enzyme mechanism. The reaction starts with a nucleophilic attack of the hydroxyl group of a tyrosine in the binding site to the phosphorus atom of a PC molecule. The result is the formation of DAG and phosphorylation of the enzyme.

The presence of a water molecule in the proximity of Tyr 223 and the phosphocholine head was also noted to aid nucleophilic substitution on phosphatidylcholine (Figure 22. c). Trajectory analyses revealed the important role of two other residues, Tyr280 and His285, in anchoring the phosphocholine head to the target in the correct position.

In isoform 2, Tyr167 plays a similar role to Tyr223 in SMS1. In the SMS2 isoform, Phe224 is responsible for anchoring the phosphocholine head. Immobilisation of the phosphocholine head allows the release of DAG and subsequent access of ceramide into the catalytic site. The two isoforms show no obvious differences in the type and duration of the interactions they establish with PC.

In both cases, the presence of tyrosine proved essential to maintain continuous contact with the phosphocholine head (Figure 22b,c). These interactions lasted for more than 90.0% of the simulation time.

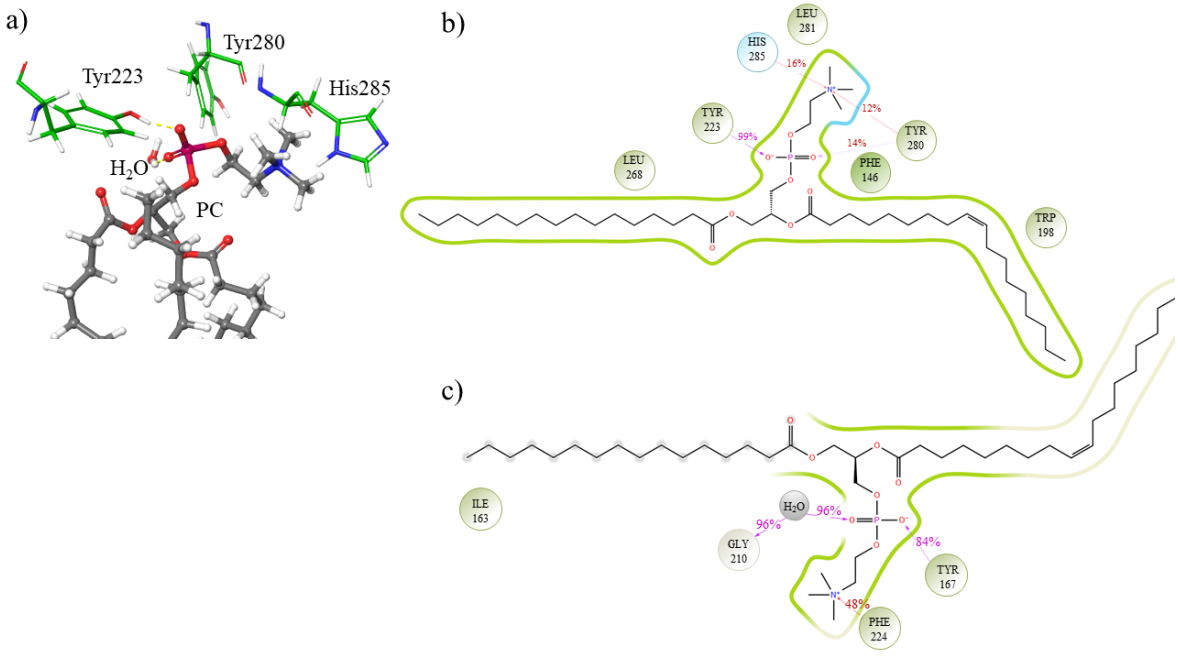

Figure 22 - A closeup of the interactions between PC and key residues at the binding site in SMS1 (a). Two-dimensional interaction map of ligand-protein contacts for SMS1/PC (b) and SMS2/PC (c).

12.4. Potential step mechanism of the reaction

All molecules involved in the conversion of PC to SM were docked with both isoforms of the receptor and, to mimic the intermediate stage described above, Tyr223 in SMS1 was modified by adding the phosphocholine group to the side chain and the same for Tyr167 in SMS2 so that the phosphorylated forms of the receptor were also obtained. Estimating the binding energy allows to reconstruct of the energy profile of the reaction. The obtained structures have been minimised, and the binding energy of each molecule has been calculated.

The mechanism of the SMS enzyme is shown in Figure 23.

PC is the natural substrate of SMS and can move freely through the membrane (1) to reach the catalytic site of the enzyme (2). The phospholipid head is transferred from the PC to a tyrosine (Tyr223 for SMS1 and Tyr167 for SMS2), leading to the formation of DAG and the phosphorylated SMS enzyme (referred to as SMS-P) (3). Subsequently, the DAG leaves the binding site and moves towards the membrane (4). The second substrate is also a component of cell membranes. Cer moves from the membrane (5) to the catalytic site of the activated form of the enzyme (6). The phosphocholine head is transferred to ceramide to form SM (7). Restoration of the non-phosphorylated form of SM is completed by moving SM into the membrane (8).

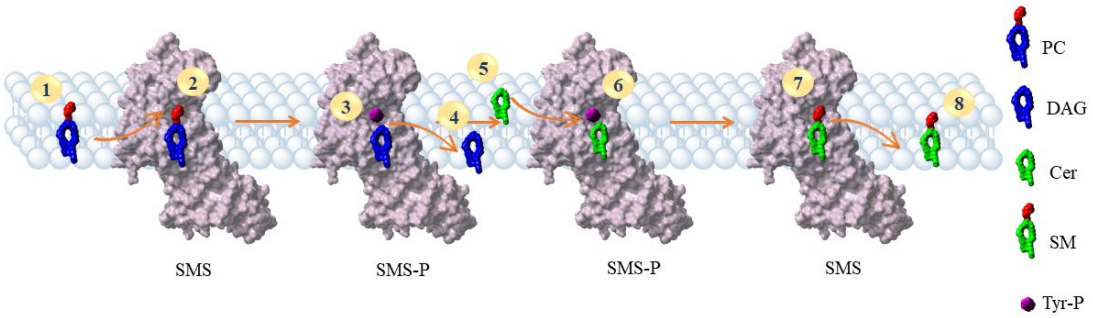

Figure 23 - Overview of the SMS reaction steps. In step (1), one of the PC molecules was highlighted. In step (2), the PC at the binding site of the SMS enzyme is observed. The transfer of the phosphocholine head to the tyrosine of the enzyme forms a modified tyrosine shown in magenta in step 3. Removing the phosphocholine head from the PC generates DAG (3). Thus, it is free to move through the membrane (4). In steps 5 and 6, ceramide moves from the membrane (5) to the SMS (6). The transfer of the phosphocholine head from the modified tyrosine to Cer generates SM (7), followed by its release into the membrane (8).

12.5. Energy profiles

After hypothesising the possible steps in the reaction, the role of hydroxylated phospholipids was investigated to better understand the function of LP561.

All intermediate stages of the possible reaction were mimicked by docking the SMS-P and SMS form with PC, Cer, DAG, SM, and the hydroxylated analogues. The binding energies of only the ligands incorporated in the PC membrane (natural and hydroxylated substrates) were

calculated to delineate the energy profile. This PC membrane was used as a model. The PC membrane used as a model consists of 120 lipids. Of these, 8 were replaced with the lipids used in the study (Figure 18) and their membrane binding energy was calculated.

To study the energy model of the reaction, three different systems were analysed. The first system represents the non-hydroxylated substrates (Figure 24a). Figure 24a shows the sum of the binding energy of PC and Cer in the membrane $(PC + Cer)_m$ (ΔG = -21.7 kcal/mol) and the sum of the binding energy of DAG and SM in the membrane $(DAG + SM)_m$ (ΔG = -26.6 kcal/mol). The $_m$ as subscript indicates the substrates immersed in the POPC model membrane.

The second system assumes the incorporation of 2-hydroxyoleic acid (2OHOA) into the PC substrate at C2 in the R configuration (2ROHPC) and in the S configuration (2SOHPC) (Figure 24b,c). The presence of the hydroxyl group at C2 of the PC leads to hydroxylated DAG (2ROHDAG and 2SOHDAG, respectively) and non-hydroxylated SM. In the third system, the inclusion of 2OHOA in the ceramide was assumed (Figure 24d,e). The hydroxylated ceramides produce two hydroxylated SM (in R and S configuration). The values of $\Delta\Delta G$ are calculated as the difference between the binding energy of the product pair minus the reactant pair immersed in a POPC membrane.

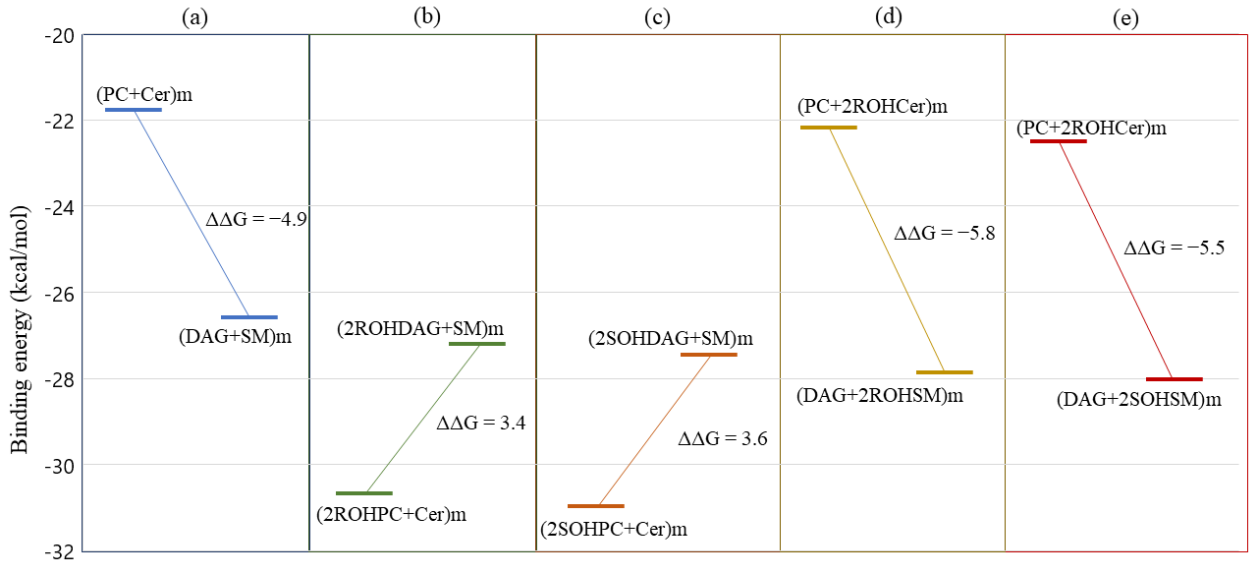

Figure 24 - Free energy differences for different hydroxylation paths in POPC membrane: non-hydroxylated system (a); hydroxylated PC with R configuration (b) and with S configuration (c); hydroxylated ceramide with R configuration (d) and S configuration (e).

It has been observed that for non-hydroxylated substrates, the equilibrium of the reaction is shifted towards DAG and SM with an energy difference ΔΔG = -4.9 kcal/mol. Hydroxylation of PC leads to less stable systems. In contrast, the presence of the hydroxyl in ceramide shifts the equilibrium towards the products by about 0.6-0.9 kcal/mol more than the non-hydroxylated forms. It has been observed that administration of 2OHOA increases SM levels in cells [129, 130]. Based on the obtained results, the increase in SM levels is consistent with the hypothesis of the incorporation of 2OHOA into ceramide.

The differences between the two isoforms (R and S) can be investigated in a similar way by calculating the binding energies in steps 2, 3, 6 and 7, of Figure 23 for natural and hydroxylated lipids. In detail, the binding energies of PC and SM in complex with SMS (SMSx/PC and SMSx/SM, respectively) and the binding energy of the phosphorylated isoform (SMSx-P) in complex with Cer and DAG (SMSx-P/Cer and SMSx-P/DAG, respectively) were compared.

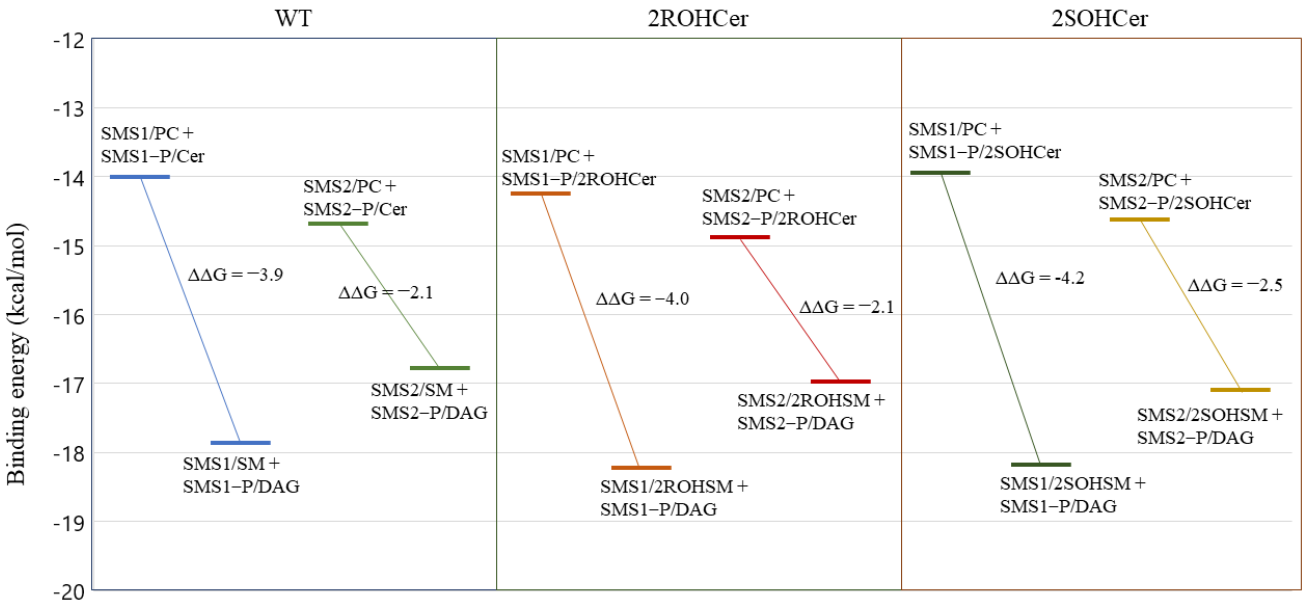

Figure 25 - Free energy of intermediate steps involving the SMS isoforms. On the left (WT) is the non-hydroxylated system; in the middle and on the right, the ceramide hydroxylated in position 2 with R configuration (2ROHCer), and S configuration (2SOHCer), respectively.

The SMS-catalysed reaction leading to the formation of SM and DAG involves several intermediate steps. First, PC has to bind to SMS (SMSx/PC), lose the phosphocholine head with the formation of a covalent bond (SMSx-P/DAG), and leave transformed into DAG. A second non-covalent binding event is then required in which Cer enters the active site (SMSx-P/Cer), acquires the choline group that was bound to the enzyme, and leaves as SM (SMSx/SM). The covalent bond between SMS and choline should not be counted because, overall, it does not contribute to the change in energy of the system.

Figure 25 shows the difference in binding energy between products and reagents in the receptor. Interestingly, the potential production of a hydroxylated SM is energetically favoured, especially for isoform 1 with a ΔΔG value of -4.0 kcal mol-1 for the R stereoisomer and -4.2 kcal mol-1 for the S stereoisomer.

The energetically favoured interaction of hydroxylated ceramides prompted us to investigate the role of hydroxylated ceramides in the transfer of the phosphocholine group from SMSx-P to ceramide. The free energy of binding (ΔG) of the molecules in complex with SMS was calculated using metadynamics.

12.6. Studying selectivity with metadynamics

Metadynamics allows a reconstruction of the free energy profile as a function of two collective variables describing the movement of ceramide (and hydroxylated analogues) in the binding site of the phosphorylated protein. Two distances were chosen as collective variables (CVs). CV1 is the distance between the phosphorus of the modified tyrosine residue (Tyr223 for SMS1 and Tyr167 for SMS2) and the oxygen atom of the ceramide (O5 atom), and CV2 is the distance between the oxygen atom in the P=O group of the modified tyrosine and the oxygen atom of the hydroxyl group of the sphingosine chain (Figure 26).

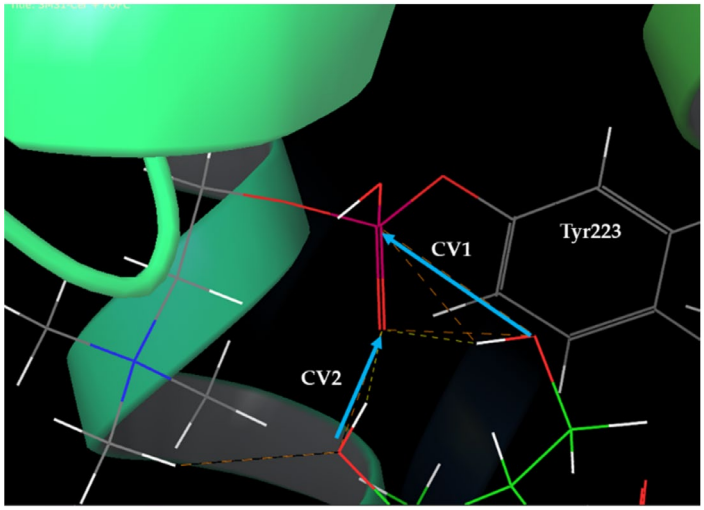

Figure 26 - Atoms chosen as CV for metadynamics simulations for the SMS1 system. The CV1 is the distance between the atom O5 of the Ceramide and the P1 of the modified tyrosine residue. The CV2 is the distance between the oxygen atom of the phosphate group and the hydroxyl group of the sphingosine chain.

The position of the minima observed for non-hydroxylated ceramide confirms the mechanism proposed for SMS1 and SMS2 (Figure 27).

The SMS1-P isoform is the enzyme with the best binding affinity for ceramide (-9.33 kcal/mol), showing two energy minima in the vicinity of the key tyrosine residue (Figure 27a). On the other hand, the SMS2-P isoform has a higher binding affinity for the hydroxylated form of the natural substrate (-9.52 kcal/mol) (Figure 27e,f).

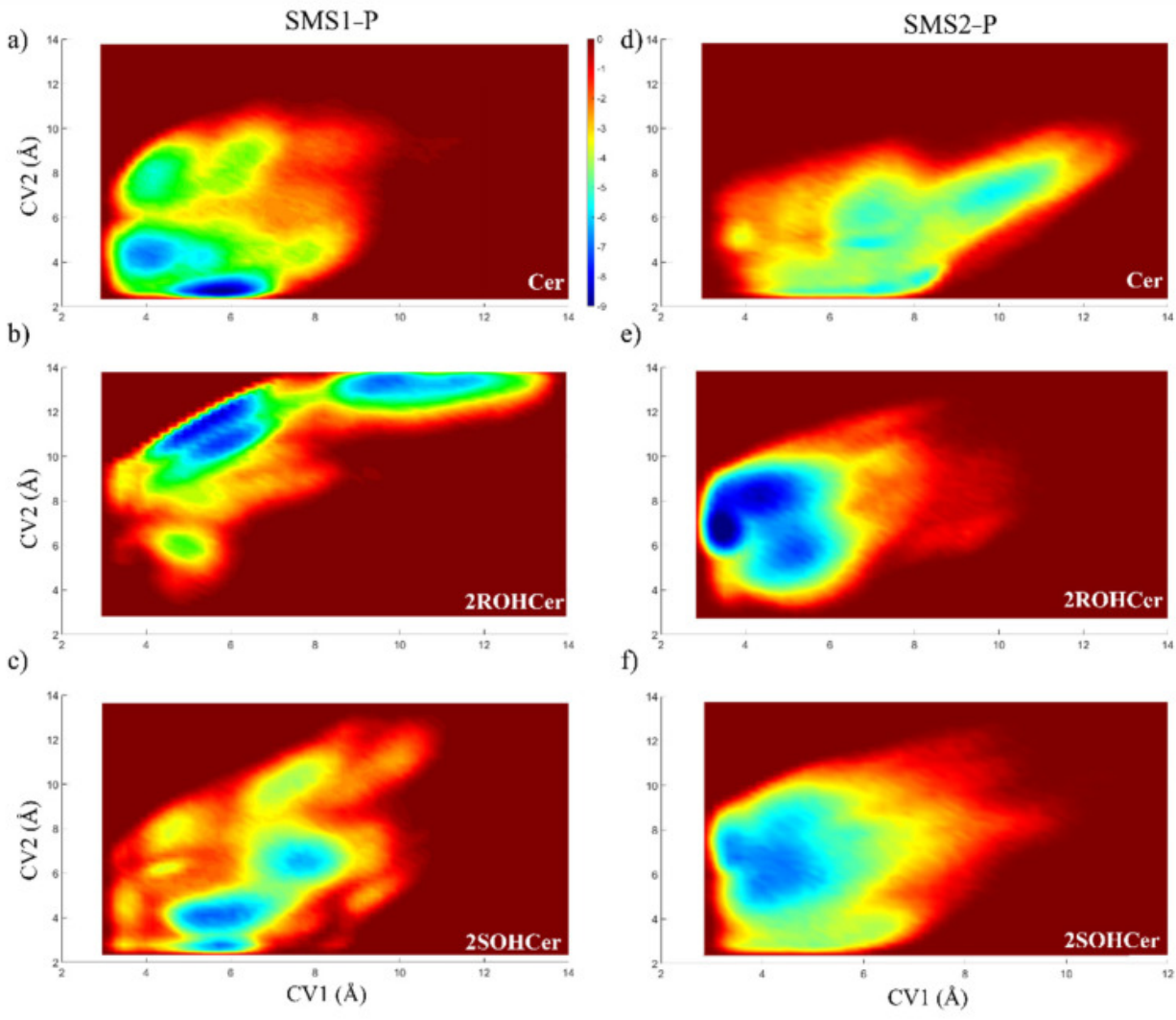

Figure 27 - Two-dimensional free-energy surface of SMS1 in complex with (a) ceramide, (b) ceramide 2R hydroxylated, (c) ceramide 2S hydroxylated; SMS2 in complex with (d) ceramide, (e) ceramide 2R hydroxylated, (f) ceramide 2S hydroxylated. The CV1 and CV2 are the x- and y-axis, respectively.

The results of the metadynamics showed the presence of a hydroxyl group allows effective binding (i.e., at a position and distance useful for transferring the choline group) to SMS2. Indeed, the presence of a hydroxyl group in the R configuration in ceramide leads to forming a hydrogen bond with the carbonyl backbone of isoleucine Ile 207 (Figure 28).

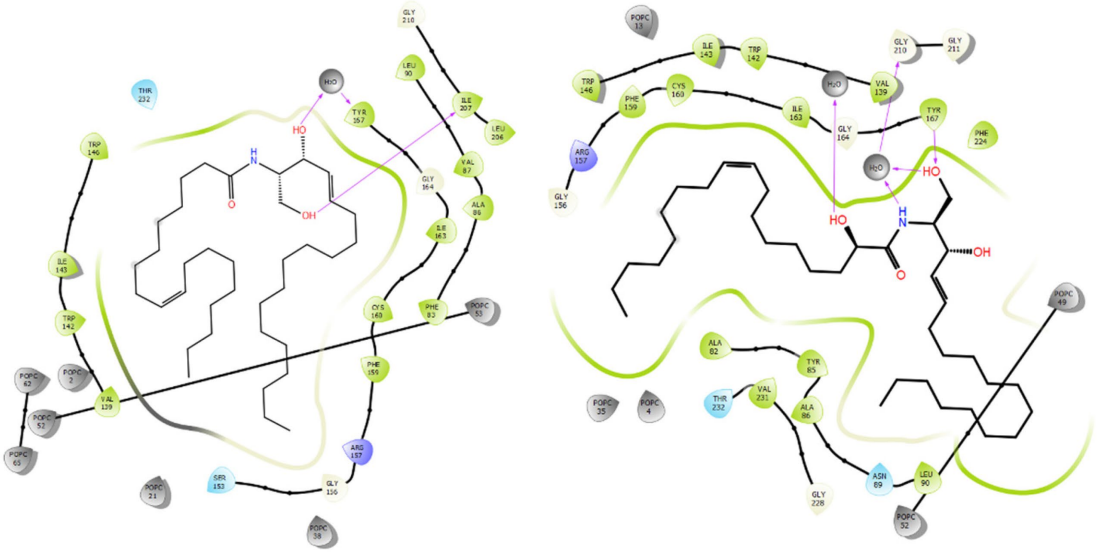

Figure 28 - The ligand interaction diagram of Ceramide (on the left) and 2ROHCer (on the right) complex with SMS2P. The pink arrows indicate the hydrogen bond that is established between the oxygen atoms of the ligand and the residues of the SMS2 protein.

Activation of SMS2, which is mainly located in the plasma membrane, results in an increase in the rate of PC→SM conversion and a modulating effect on the physical state of the plasma membrane. It is well known that the composition of the plasma membrane plays a key role in controlling the function of numerous proteins [131], and these results could clarify the molecular mechanism of LP561 in increasing SM levels.

PART II CONCLUSIONS

13. Conclusions

Sphingolipids are key molecules in the regulation of the cell cycle, apoptosis, angiogenesis, stress, and inflammatory responses. It is known in the literature that patients with glioblastoma show reductions in the level of sphingomyelin (SM) in plasma membranes, and results from early clinical trials have shown that taking LP561 (2OHOA) restores SM values. SM is produced by SMS (sphingomyelin synthase). The 3D structures of the two SMS isoforms were predicted and subsequently used for further computational investigations. The binding site of both SMS1 and SMS2 was identified in the transmembrane region by using a new plugin that exploits the movement of water molecules in the system and the identification of possible conservation sites.

Analysis of MD trajectories revealed a key role for tyrosine in the binding site; they are involved in the mechanism of both SMS isoforms (Tyr223 for SMS1 and Tyr167 for SMS2). The MD also showed a water molecule in the catalytic area, which helps the nucleophilic attachment of the phosphocholine head to the receptor. In this way, it was possible to determine the energy profile of the PC→SM transformation and define the intermediate steps in both SMS1 and SMS2.

An interesting difference between SMS1 and SMS2 towards the hydroxylated species appeared from the metadynamics results. It was shown that hydroxylated ceramide, especially in the R configuration, is largely favoured in the interaction with SMS2. Since SMS2 is mainly localised in the plasma membrane, these results suggest a possible role for some hydroxylated species in the homeostasis of lipid composition. The possible role of 2OHOA and the action of SMS was therefore clarified.

In the future, it is planned to extend the study to other receptors and lipid species. Indeed, data in the literature confirm the production of 2-hydroxylated acyl chains by the stereospecific

enzyme FA2H. The obtained results suggest a possible role for another enzyme, ceramide synthase, in incorporating 2OHOA (particularly in the R form) into ceramide.

REFERENCES

1. Kuntz, I.D., et al., *A geometric approach to macromolecule-ligand interactions.* Journal of molecular biology, 1982. **161**(2): p. 269-288.
2. Van Zundert, G., et al., *The HADDOCK2. 2 web server: user-friendly integrative modeling of biomolecular complexes.* Journal of molecular biology, 2016. **428**(4): p. 720-725.
3. Brown, F.K., et al., *The evolution of drug design at Merck Research Laboratories.* Journal of computer-aided molecular design, 2017. **31**(3): p. 255-266.
4. Canese, K. and S. Weis, *PubMed: the bibliographic database.* The NCBI Handbook, 2013. **2**: p. 1.
5. Lopes, P.E., O. Guvench, and A.D. MacKerell, *Current status of protein force fields for molecular dynamics simulations*, in *Molecular modeling of proteins*. 2015, Springer. p. 47-71.
6. Monticelli, L. and D.P. Tieleman, *Force fields for classical molecular dynamics.* Biomolecular simulations, 2013: p. 197-213.
7. Fischer, E., *Einfluss der Configuration auf die Wirkung der Enzyme.* Berichte der deutschen chemischen Gesellschaft, 1894. **27**(3): p. 2985-2993.
8. Dastmalchi, S., *Methods and algorithms for molecular docking-based drug design and discovery*. 2016: IGI Global.
9. Lengauer, T. and M. Rarey, *Computational methods for biomolecular docking.* Current opinion in structural biology, 1996. **6**(3): p. 402-406.
10. Vajda, S. and F. Guarnieri, *Characterization of protein-ligand interaction sites using experimental and computational methods.* Current opinion in drug discovery & development, 2006. **9**(3): p. 354-362.
11. Sousa, S.F., P.A. Fernandes, and M.J. Ramos, *Protein–ligand docking: current status and future challenges.* Proteins: Structure, Function, and Bioinformatics, 2006. **65**(1): p. 15-26.
12. Huang, S.-Y. and X. Zou, *Advances and challenges in protein-ligand docking.* International journal of molecular sciences, 2010. **11**(8): p. 3016-3034.
13. Spyrakis, F., P. Cozzini, and G.E. Kellogg, *Docking and scoring in drug discovery.* Burger's Medicinal Chemistry and Drug Discovery, 2003: p. 601-684.
14. Kleywegt, G.J. and T.A. Jones, *Detection, delineation, measurement and display of cavities in macromolecular structures.* Acta Crystallographica Section D: Biological Crystallography, 1994. **50**(2): p. 178-185.
15. Campbell, S.J., et al., *Ligand binding: functional site location, similarity and docking.* Current opinion in structural biology, 2003. **13**(3): p. 389-395.
16. Tripathi, A. and G.E. Kellogg, *A novel and efficient tool for locating and characterizing protein cavities and binding sites.* Proteins: Structure, Function, and Bioinformatics, 2010. **78**(4): p. 825-842.
17. Leach, A.R., *Ligand docking to proteins with discrete side-chain flexibility.* Journal of molecular biology, 1994. **235**(1): p. 345-356.
18. Cozzini, P., et al., *Target flexibility: an emerging consideration in drug discovery and design.* Journal of medicinal chemistry, 2008. **51**(20): p. 6237-6255.

19. Heaslet, H., et al., *Conformational flexibility in the flap domains of ligand-free HIV protease.* Acta Crystallographica Section D: Biological Crystallography, 2007. **63**(8): p. 866-875.
20. Latorraca, N.R., A. Venkatakrishnan, and R.O. Dror, *GPCR dynamics: structures in motion.* Chemical reviews, 2017. **117**(1): p. 139-155.
21. Muegge, I. and M. Rarey, *Small molecule docking and scoring.* Reviews in computational chemistry, 2001. **17**: p. 1-60.
22. Yadava, U., *Search algorithms and scoring methods in protein-ligand docking.* Endocrinol Int J, 2018. **6**(6): p. 359-367.
23. LEACH, A., *Principle and Applications of Molecular Modeling*. 1996, Addison Wesley Longman Limited, Harlow, England.
24. Taylor, R.D., P.J. Jewsbury, and J.W. Essex, *A review of protein-small molecule docking methods.* Journal of computer-aided molecular design, 2002. **16**(3): p. 151-166.
25. Norberg, J. and L. Nilsson, *Advances in biomolecular simulations: methodology and recent applications.* Quarterly Reviews of Biophysics, 2003. **36**(3): p. 257-306.
26. Sousa, S.F., et al., *Protein-ligand docking in the new millennium–a retrospective of 10 years in the field.* Current medicinal chemistry, 2013. **20**(18): p. 2296-2314.
27. Kitchen, D.B., et al., *Docking and scoring in virtual screening for drug discovery: methods and applications.* Nature reviews Drug discovery, 2004. **3**(11): p. 935-949.
28. Dinur, U. and A.T. Hagler, *New approaches to empirical force fields.* Reviews in computational chemistry, 1991. **2**: p. 99-164.
29. Koppensteiner, W. and M.J. Sippl, *Knowledge-based potentials--back to the roots.* Biochemistry. Biokhimiia, 1998. **63**(3): p. 247-252.
30. Liu, Q., C.K. Kwoh, and J. Li, *Binding affinity prediction for protein–ligand complexes based on β contacts and B factor.* Journal of chemical information and modeling, 2013. **53**(11): p. 3076-3085.
31. Meng, X.-Y., et al., *Molecular docking: a powerful approach for structure-based drug discovery.* Current computer-aided drug design, 2011. **7**(2): p. 146-157.
32. Debroise, T., E.I. Shakhnovich, and N. Chéron, *A hybrid knowledge-based and empirical scoring function for protein–ligand interaction: SMoG2016.* Journal of chemical information and modeling, 2017. **57**(3): p. 584-593.
33. Klebe, G., *Recent developments in structure-based drug design.* Journal of Molecular Medicine, 2000. **78**(5): p. 269-281.
34. Ghasemi, J.B., A. Abdolmaleki, and F. Shiri, *Molecular docking challenges and limitations*, in *Pharmaceutical sciences: Breakthroughs in research and practice*. 2017, IGI Global. p. 770-794.
35. Chen, Y.-C., *Beware of docking!* Trends in pharmacological sciences, 2015. **36**(2): p. 78-95.
36. Breiten, B., et al., *Water networks contribute to enthalpy/entropy compensation in protein–ligand binding.* Journal of the American Chemical Society, 2013. **135**(41): p. 15579-15584.
37. Hamelberg, D. and J.A. McCammon, *Standard free energy of releasing a localized water molecule from the binding pockets of proteins: double-decoupling method.* Journal of the American Chemical Society, 2004. **126**(24): p. 7683-7689.
38. Teague, S.J., *Implications of protein flexibility for drug discovery.* Nature reviews Drug discovery, 2003. **2**(7): p. 527-541.
39. Erickson, J.A., et al., *Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy.* Journal of medicinal chemistry, 2004. **47**(1): p. 45-55.

40. Österberg, F., et al., *Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock.* Proteins: Structure, Function, and Bioinformatics, 2002. **46**(1): p. 34-40.
41. Hart, T.N. and R.J. Read, *A multiple-start Monte Carlo docking method.* Proteins: Structure, Function, and Bioinformatics, 1992. **13**(3): p. 206-222.
42. Morris, G.M., et al., *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.* Journal of computational chemistry, 1998. **19**(14): p. 1639-1662.
43. Knegtel, R.M., I.D. Kuntz, and C. Oshiro, *Molecular docking to ensembles of protein structures.* Journal of molecular biology, 1997. **266**(2): p. 424-440.
44. Oshiro, C.M., I.D. Kuntz, and J.S. Dixon, *Flexible ligand docking using a genetic algorithm.* Journal of computer-aided molecular design, 1995. **9**(2): p. 113-130.
45. Apostolakis, J., A. Plückthun, and A. Caflisch, *Docking small ligands in flexible binding sites.* Journal of Computational Chemistry, 1998. **19**(1): p. 21-37.
46. Pak, Y. and S. Wang, *Application of a molecular dynamics simulation method with a generalized effective potential to the flexible molecular docking problems.* The Journal of Physical Chemistry B, 2000. **104**(2): p. 354-359.
47. Schnecke, V., et al., *Screening a peptidyl database for potential ligands to proteins with side-chain flexibility.* Proteins: Structure, Function, and Bioinformatics, 1998. **33**(1): p. 74-87.
48. Carlson, H.A., *Protein flexibility is an important component of structure-based drug discovery.* Current Pharmaceutical Design, 2002. **8**(17): p. 1571-1578.
49. Barril, X. and S.D. Morley, *Unveiling the full potential of flexible receptor docking using multiple crystallographic structures.* Journal of medicinal chemistry, 2005. **48**(13): p. 4432-4443.
50. Ahmed, M.H., et al., *Understanding Water and Its Many Roles in Biological Structure: Ways to Exploit a Resource for Drug Discovery*, in *Computer-Aided Drug Discovery*. 2015, Springer. p. 85-110.
51. Biela, A., et al., *Dissecting the hydrophobic effect on the molecular level: the role of water, enthalpy, and entropy in ligand binding to thermolysin.* Angewandte Chemie International Edition, 2013. **52**(6): p. 1822-1828.
52. Krimmer, S.G., et al., *Rational design of thermodynamic and kinetic binding profiles by optimizing surface water networks coating protein-bound ligands.* Journal of medicinal chemistry, 2016. **59**(23): p. 10530-10548.
53. Fornabaio, M., et al., *Simple, intuitive calculations of free energy of binding for protein− ligand complexes. 3. The free energy contribution of structural water molecules in HIV-1 protease complexes.* Journal of medicinal chemistry, 2004. **47**(18): p. 4507-4516.
54. Finkelstein, A.V. and J. Janin, *The price of lost freedom: entropy of bimolecular complex formation.* Protein Engineering, Design and Selection, 1989. **3**(1): p. 1-3.
55. Salaniwal, S., et al., *Critical evaluation of methods to incorporate entropy loss upon binding in high-throughput docking.* Proteins: Structure, Function, and Bioinformatics, 2007. **66**(2): p. 422-435.
56. Abel, R., et al., *Role of the active-site solvent in the thermodynamics of factor Xa ligand binding.* Journal of the American Chemical Society, 2008. **130**(9): p. 2817-2831.
57. Berman, H.M., et al., *The protein data bank.* Nucleic acids research, 2000. **28**(1): p. 235-242.

58. Liu, T., et al., *BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities.* Nucleic acids research, 2007. **35**(suppl_1): p. D198-D201.
59. Krieger, E. and G. Vriend, *New ways to boost molecular dynamics simulations.* Journal of computational chemistry, 2015. **36**(13): p. 996-1007.
60. BIOVIA, D.S., Materials Studio, 17.1.0.48, San Diego: Dassault Systèmes, 2017.
61. Piotto, S., et al., *Yada: a novel tool for molecular docking calculations.* Journal of computer-aided molecular design, 2016. **30**(9): p. 753-759.
62. Van Meir, E.G., et al., *Exciting new advances in neuro-oncology: the avenue to a cure for malignant glioma.* CA: a cancer journal for clinicians, 2010. **60**(3): p. 166-193.
63. Hadjipanayis, C.G. and E.G. Van Meir, *Brain cancer propagating cells: biology, genetics and targeted therapies.* Trends in molecular medicine, 2009. **15**(11): p. 519-530.
64. Furnari, F.B., et al., *Malignant astrocytic glioma: genetics, biology, and paths to treatment.* Genes & development, 2007. **21**(21): p. 2683-2710.
65. Stiver, S., et al., *VEGF-A angiogenesis induces a stable neovasculature in adult murine brain.* Journal of Neuropathology & Experimental Neurology, 2004. **63**(8): p. 841-855.
66. Carlsson, S.K., S.P. Brothers, and C. Wahlestedt, *Emerging treatment strategies for glioblastoma multiforme.* EMBO molecular medicine, 2014. **6**(11): p. 1359-1370.
67. Escriba, P.V., P. Morales, and A. Smith, *Membrane Phospholipid Reorganization Differentially Regulates Metallothionein and Heme Oxygenase by Heme–Hemopexin.* DNA and cell biology, 2002. **21**(4): p. 355-364.
68. Chen, N.-T., et al., *Probing the dynamics of doxorubicin-DNA intercalation during the initial activation of apoptosis by fluorescence lifetime imaging microscopy (FLIM).* 2012.
69. Lüpertz, R., et al., *Dose-and time-dependent effects of doxorubicin on cytotoxicity, cell cycle and apoptotic cell death in human colon cancer cells.* Toxicology, 2010. **271**(3): p. 115-121.
70. Bhagat, A. and E.S. Kleinerman, *Anthracycline-induced cardiotoxicity: causes, mechanisms, and prevention.* Current Advances in Osteosarcoma, 2020: p. 181-192.
71. Doroshow, J.H., *Mechanisms of anthracycline-enhanced reactive oxygen metabolism in tumor cells.* Oxidative medicine and cellular longevity, 2019. **2019**.
72. Beretta, G.L. and F. Zunino, *Molecular mechanisms of anthracycline activity*, in *Anthracycline chemistry and biology II*. 2007, Springer. p. 1-19.
73. Fujiwara, A., T. Hoshino, and J.W. Westley, *Anthracycline antibiotics.* Critical Reviews in Biotechnology, 1985. **3**(2): p. 133-157.
74. Krohn, K., *Anthracycline chemistry and biology I: biological occurence and biosynthesis, synthesis and chemistry*. Vol. 282. 2009: Springer.
75. Bachur, N.R., *Anthracycline antibiotic pharmacology and metabolism.* Cancer Treat Rep, 1979. **63**(8): p. 7-820.
76. McGowan, J.V., et al., *Anthracycline chemotherapy and cardiotoxicity.* Cardiovascular drugs and therapy, 2017. **31**(1): p. 63-75.
77. Wirtz, V.J., et al., *Essential medicines for universal health coverage.* The Lancet, 2017. **389**(10067): p. 403-476.
78. Minotti, G., et al., *Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity.* Pharmacological reviews, 2004. **56**(2): p. 185-229.
79. Weiss, R.B. *The anthracyclines: will we ever find a better doxorubicin?* in *Seminars in oncology*. 1992.

80. Von Holst, H., et al., *Uptake of adriamycin in tumour and surrounding brain tissue in patients with malignant gliomas.* Acta neurochirurgica, 1990. **104**(1): p. 13-16.
81. da Ros, M., et al., *The use of anthracyclines for therapy of CNS tumors.* Anti-cancer agents in medicinal chemistry, 2015. **15**(6): p. 721.
82. Escriba, P.V., M. Sastre, and J.A. Garcia-Sevilla, *Disruption of cellular signaling pathways by daunomycin through destabilization of nonlamellar membrane structures.* Proceedings of the National Academy of Sciences, 1995. **92**(16): p. 7595-7599.
83. Lladó, V., et al., *Pivotal role of dihydrofolate reductase knockdown in the anticancer activity of 2-hydroxyoleic acid.* Proceedings of the National Academy of Sciences, 2009. **106**(33): p. 13754-13758.
84. Llado, V., et al., *Minerval induces apoptosis in Jurkat and other cancer cells.* Journal of cellular and molecular medicine, 2010. **14**(3): p. 659-670.
85. Terés, S., et al., *2-Hydroxyoleate, a nontoxic membrane binding anticancer drug, induces glioma cell differentiation and autophagy.* Proceedings of the National Academy of Sciences, 2012. **109**(22): p. 8489-8494.
86. Barceló, F., et al., *The hypotensive drug 2-hydroxyoleic acid modifies the structural properties of model membranes.* Molecular membrane biology, 2004. **21**(4): p. 261-268.
87. Martínez, J., et al., *Membrane structure modulation, protein kinase Cα activation, and anticancer activity of minerval.* Molecular pharmacology, 2005. **67**(2): p. 531-540.
88. Gajate, C. and F. Mollinedo, *Edelfosine and perifosine induce selective apoptosis in multiple myeloma by recruitment of death receptors and downstream signaling molecules into lipid rafts.* Blood, 2007. **109**(2): p. 711-719.
89. Bartke, N. and Y.A. Hannun, *Bioactive sphingolipids: metabolism and function.* Journal of lipid research, 2009. **50**: p. S91-S96.
90. Simons, K. and D. Toomre, *Lipid rafts and signal transduction.* Nature reviews Molecular cell biology, 2000. **1**(1): p. 31-39.
91. De Almeida, R.F., A. Fedorov, and M. Prieto, *Sphingomyelin/phosphatidylcholine/cholesterol phase diagram: boundaries and composition of lipid rafts.* Biophysical journal, 2003. **85**(4): p. 2406-2416.
92. Hannun, Y.A. and L.M. Obeid, *Principles of bioactive lipid signalling: lessons from sphingolipids.* Nature reviews Molecular cell biology, 2008. **9**(2): p. 139-150.
93. Huitema, K., et al., *Identification of a family of animal sphingomyelin synthases.* The EMBO journal, 2004. **23**(1): p. 33-44.
94. Piotto, S., et al., *Computational study on human sphingomyelin synthase 1 (hSMS1).* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2017. **1859**(9): p. 1517-1525.
95. Holthuis, J.C. and C. Luberto, *Tales and mysteries of the enigmatic sphingomyelin synthase family.* Sphingolipids as Signaling and Regulatory Molecules, 2010: p. 72-85.
96. Villani, M., et al., *Sphingomyelin synthases regulate production of diacylglycerol at the Golgi.* Biochemical Journal, 2008. **414**(1): p. 31-41.
97. Li, Z., et al., *Inhibition of sphingomyelin synthase (SMS) affects intracellular sphingomyelin accumulation and plasma membrane lipid organization.* Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 2007. **1771**(9): p. 1186-1194.
98. Tafesse, F.G., P. Ternes, and J.C. Holthuis, *The multigenic sphingomyelin synthase family.* Journal of Biological Chemistry, 2006. **281**(40): p. 29421-29425.
99. Adada, M., C. Luberto, and D. Canals, *Inhibitors of the sphingomyelin cycle: Sphingomyelin synthases and sphingomyelinases.* Chemistry and physics of lipids, 2016. **197**: p. 45-59.

100. Lopez, D.H., et al., *2-Hydroxy arachidonic acid: a new non-steroidal anti-inflammatory drug.* PloS one, 2013. **8**(8): p. e72052.
101. Casas, J., et al., *G protein-membrane interactions II: effect of G protein-linked lipids on membrane structure and G protein-membrane interactions.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2017. **1859**(9): p. 1526-1535.
102. Marques, J.T., H.S. Marinho, and R.F. de Almeida, *Sphingolipid hydroxylation in mammals, yeast and plants–an integrated view.* Progress in lipid research, 2018. **71**: p. 18-42.
103. Jaikishan, S. and J.P. Slotte, *Stabilization of sphingomyelin interactions by interfacial hydroxyls—A study of phytosphingomyelin properties.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2013. **1828**(2): p. 391-397.
104. Piotto, S., et al., *The effect of hydroxylated fatty acid-containing phospholipids in the remodeling of lipid membranes.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2014. **1838**(6): p. 1509-1517.
105. Hama, H., *Fatty acid 2-Hydroxylation in mammalian sphingolipid biology.* Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 2010. **1801**(4): p. 405-414.
106. Herrero, A.B., et al., *Levels of SCS7/FA2H-mediated fatty acid 2-hydroxylation determine the sensitivity of cells to antitumor PM02734.* Cancer research, 2008. **68**(23): p. 9779-9787.
107. Guo, L., et al., *Stereospecificity of fatty acid 2-hydroxylase and differential functions of 2-hydroxy fatty acid enantiomers.* Journal of lipid research, 2012. **53**(7): p. 1327-1335.
108. Mizutani, Y., et al., *2-Hydroxy-ceramide synthesis by ceramide synthase family: enzymatic basis for the preference of FA chain length.* Journal of lipid research, 2008. **49**(11): p. 2356-2364.
109. Schaeren-Wiemers, N., P. Van der Bijl, and M. Schwab, *The UDP-galactose: ceramide galactosyltransferase: Expression pattern in oligodendrocytes and Schwann cells during myelination and substrate preference for hydroxyceramide.* Journal of neurochemistry, 1995. **65**(5): p. 2267-2278.
110. *UniProt: the universal protein knowledgebase.* Nucleic acids research, 2017. **45**(D1): p. D158-D169.
111. Madeira, F., et al., *The EMBL-EBI search and sequence analysis tools APIs in 2019.* Nucleic acids research, 2019. **47**(W1): p. W636-W641.
112. https://smp.softmining.it, S.P.P.b.S.s.A.o.
113. Laskowski, R.A., et al., *PROCHECK: a program to check the stereochemical quality of protein structures.* Journal of applied crystallography, 1993. **26**(2): p. 283-291.
114. Ashkenazy, H., et al., *ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids.* Nucleic acids research, 2010. **38**(suppl_2): p. W529-W533.
115. Debiec, K.T., et al., *Further along the road less traveled: AMBER ff15ipq, an original protein force field built on a self-consistent physical model.* Journal of chemical theory and computation, 2016. **12**(8): p. 3926-3947.
116. Miyamoto, S. and P.A. Kollman, *Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models.* Journal of computational chemistry, 1992. **13**(8): p. 952-962.
117. Maier, J.A., et al., *ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB.* Journal of chemical theory and computation, 2015. **11**(8): p. 3696-3713.

118. Sessa, L., et al. *A new flexible protocol for docking studies*. in *Italian Workshop on Artificial Life and Evolutionary Computation*. 2015. Springer.
119. Sessa, L. and S. Piotto, *Molecular Dynamics and Morphing Protocols for High Accuracy Molecular Docking*, in *Advances in Bionanomaterials*. 2018, Springer. p. 85-96.
120. Bowers, K.J., et al. *Scalable algorithms for molecular dynamics simulations on commodity clusters*. in *SC'06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*. 2006. IEEE.
121. Roos, K., et al., *OPLS3e: Extending force field coverage for drug-like small molecules.* Journal of chemical theory and computation, 2019. **15**(3): p. 1863-1874.
122. Lomize, M.A., et al., *OPM database and PPM web server: resources for positioning of proteins in membranes.* Nucleic acids research, 2012. **40**(D1): p. D370-D376.
123. Schrödinger Release 2021-4: Maestro, S., LLC, New York, NY, 2021.
124. Käll, L., A. Krogh, and E.L. Sonnhammer, *Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server.* Nucleic acids research, 2007. **35**(suppl_2): p. W429-W432.
125. de Vries, S.J., A.D. van Dijk, and A.M. Bonvin, *WHISCY: what information does surface conservation yield? Application to data-driven docking.* Proteins: Structure, Function, and Bioinformatics, 2006. **63**(3): p. 479-489.
126. Ouzounis, C., et al. *Are binding residues conserved?* in *Pacific symposium on biocomputing. Pacific symposium on biocomputing*. 1998.
127. Celniker, G., et al., *ConSurf: using evolutionary data to raise testable hypotheses about protein function.* Israel Journal of Chemistry, 2013. **53**(3-4): p. 199-206.
128. Yeang, C., et al., *The domain responsible for sphingomyelin synthase (SMS) activity.* Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 2008. **1781**(10): p. 610-617.
129. Lou, B., et al., *2-Hydroxy-oleic acid does not activate sphingomyelin synthase activity.* Journal of Biological Chemistry, 2018. **293**(47): p. 18328-18336.
130. Barceló-Coblijn, G., et al., *Sphingomyelin and sphingomyelin synthase (SMS) in the malignant transformation of glioma cells and in 2-hydroxyoleic acid therapy.* Proceedings of the National Academy of Sciences, 2011. **108**(49): p. 19569-19574.
131. Török, Z., et al., *Plasma membranes as heat stress sensors: from lipid-controlled molecular switches to therapeutic applications.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2014. **1838**(6): p. 1594-1618.

LIST OF ABBREVIATIONS

ΔG	Free energy of binding
2OHFA	2-hydroxylated fatty acids
2OHOA	2-hydroxyoleic acid
AUC	Area Under the Curve
CADD	Computer aided drug design
CDKs	Cyclin-dependent kinases
Cer	Ceramide
CerS	Ceramide synthases
CNS	Central nervous system
CV	Collective variable
DAG	Diacylglycerol
EMA	European Medicines Agency
FA2H	Fatty acid 2-hydroxylase
FPR	False positive rate
GA	Genetic Algorithm
GBM	Glioblastoma
GFA	Genetic Function Algorithm
IDH	Isocitrate dehydrogenases
LPP	Lipid phosphatase family
MAPK	Mitogen-activated protein kinase
MC	Monte Carlo
MD	Molecular Dynamic
NPT	Normal pressure and temperature
PC	Phosphatidylcholine
PDB	Protein Data Bank
PKC	Protein kinase C
PME	Particle Mesh Edwald
R^2	Coefficient of determination
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square fluctuation
ROC	Receiver Operating Characteris

SAM	Sterile Alpha Motif
SM	Sphingomyelin
SMS	Sphingomyelin Synthase
SoC	Standard of Care
TMZ	Temozolomide
TRP	True positive rate
V_{H2O}	Variability of hydrogen bonds
WHO	World Health Organisation