POLITECNICO DI TORINO

Master Degree in Biomedical Engineering
Master Thesis

Fragmented Molecular Docking to Rationally Improve the Accuracy of Blind Ligand-Receptor Binding Prediction

Thesis Advisor:
Prof. M. A. Deriu

Co-Advisor:
Dr. G. Grasso

Candidate:
Arianna Di Gregorio

ACADEMIC YEAR 2019-2020

Abstract

Molecular docking is a computational screening approach in drug design able to predict the conformation of a protein-ligand complex. Docking algorithms provide an efficient and cost-effective support to experimental techniques and high-throughput screening. Molecular docking is generally applied starting from the knowledge of the protein binding region. However, a precise information about the correct binding site is often missing and it becomes necessary to explore the entire protein surface by docking algorithms. In the present thesis, a new methodology to identify the experimental binding pose of small molecule ligands into protein structures where the real binding sites are unknown will be presented. The approach consists of carrying out ligand-protein docking separately in multiple fragmented boxes, shifting the location of the box step by step, in order to cover the entire surface of the protein. This fragmented docking has been compared with the blind docking performed by standard docking protocols on 116 protein-ligand complexes of Heat Shock Protein 90 – alpha and 176 of Human Immunodeficiency virus protease 1. The fragmented docking has demonstrated its ability to identify more accurate docking poses than blind docking performed by AutoDock-Vina. In order to improve the docking results Molecular-Mechanics/Generalized-Born-Surface-Area has been employed to re-score the docking outcomes. The results deriving from this rescoring show that MM/GBSA is able further increase the accuracy of the approach. The method is relived a good compromise between accuracy and computational effort. Future studies are needed to overcome the main limitation of the present algorithm, which is related to the conformational plasticity of the protein targets.

1 Introduction

The current chapter introduces the present Master Thesis research, elucidating aims and objectives.

Aim of the thesis

The aim of the thesis is to develop a new methodology of protein-ligand docking able to improve the experimental binding modes and affinities of small molecules within the binding site of particular receptor targets when the binding site in unknown. Protein-ligand docking is the procedure performed to predict the position and orientation of a ligand when it is bound to a receptor.

Organization of the thesis

The present thesis is divided in sections briefly described as following:

Chapter 1 is the present introductory part.

Chapter 2 is devoted to illustrating the process leading to the discovery and development of new drugs. Since the affinity between a target and a drug has a crucial importance in order to find efficacy in vivo drugs, an explanation of physicochemical mechanisms underlying protein-ligand binding is provided. Finally, the methods available for investigating protein–ligand binding affinity, including experimental and computational approaches are presented.

Chapter 3 provides a theoretical overview of the methods employed in the present work. Concept of molecular modelling for investigating biological mechanisms is introduced. Then, molecular docking is discussed, focusing on search algorithm and energy scoring function for generating and evaluating ligand poses. Finally, Molecular Mechanics approach is introduced to provide a background on physical basis behind molecular modelling.

Chapter 4 presents a novel protocol used to perform protein-ligand docking. Furthermore, the top discovered hits have been reported with a detailed discussion about the validity of the approach.

Chapter 5 is devoted to general conclusions.

2 Rational Drug Discovery

This chapter is devoted to illustrating the process leading to the discovery and development of new drugs. Since the affinity between a target and a drug has a crucial importance in order to find efficacy in vivo drugs, an explanation of physicochemical mechanisms underlying protein-ligand binding is provided. Finally, the methods available for investigating protein–ligand binding affinity, including experimental and computational approaches are presented.

2.1 Drug Discovery Pipeline

Discovery and development of a new drug is generally known as a very complex process which requires a lot of time and resources. It has been estimated that each new drug employs 14 years to develop, costing about $800 million. In Figure 1 is presented the drug discovery process.
The starting point is to select a relevant target. Target discovery is composed of three steps: the provision of disease models, target identification and target validation[1]. Target identification and validation can be achieved whit molecular and system approach. The molecular strategy employs techniques such as genomics, proteomics, genetic association, forward genetics and reverse genetics, whereas the systems strategy employs clinical and in vivo studies to identify potential targets. A genomics approach tries to identify the disease targets at the level of gene expression through the comparison of normal and diseased tissue. Proteomics measures protein expression, activity and interaction with other biological macromolecules in order to understand cellular function. Genetic association identifies relationship between mutations in genes in order to determine disease mechanism to target identification. Forward genetics seeks to find the genetic basis of a phenotype using cell and animal models, while reverse genetics proceeds in the opposite direction by analyzing the phenotypic effects of specific gene sequences obtained by DNA sequencing to discovering the function of a gene. These experimental methods are laborious and time consuming, then a series of computational tools have also been developed in order to improve target identification. Computational tools can be categorized into sequence-based approach and structure-based approach. Sequence-based methods include sequence alignment for gene selection, prioritization of protein families, gene and protein annotation, and expression data analysis for microarray or gene chip. A method employed in structure-based is reverse docking, which consists of docking a compound with certain biological activities in the binding sites of all three-dimensional structures in a given protein database. The advantage of

reverse docking is in addition to identifying target candidates for active compounds, it is also possible to identify potential targets responsible for the toxicity or side effects of a drug. However, reverse docking still has some limitations because the proteins present in the Protein Data Bank[2] (PDB) databases are not sufficient to cover all the protein information of disease and the approach does not consider protein flexibility during docking simulation. These two aspects and the inaccuracy of the scoring functions for reverse docking will produce false negatives.

After identifying the target, it must be validated to demonstrate the functional role of the potential target in the disease phenotype. Target need to be considered druggable. It means that the protein should have a binding site which can contain a drug-like compound with sufficient affinity and specificity.

The second step is hit generation. A hit is a compound that binds to the target and has the desired effect. In order to identify hits, it is needed screening a compound collection on the selected target. The compound collection consists of natural or synthesized products. The screening can be achieved with experimental or computational methods. High throughput screening (HTS) is an experimental method that involves screening the entire library of compounds on the target in complex laboratories and without a priori knowledge on the nature of the chemotype that could have activity on the target protein. A computational method is virtual screening that screens the large number of compounds on a target and measure compounds that exhibit activity at a set concentration. The hit identification phase lasts around 6 months. The output of hit identification is a set of compounds whose chemical structures have been checked and which have reproducibility been shown to have activity.

Before the lead optimization, the hits to leads phase establishes which compounds has the potential to be optimized into a drug candidate. According to the Lipinski Rule of Five[3], a molecule to became drug should to have no more than 5 hydrogen bond donors, 10 hydrogen bond acceptors, a molecular weight less than 500 Daltons and the octanol-water partition coefficient (LogP) lower than 5. In this phase significant resources are spent in optimizing the properties of compounds which are re-synthesized. The aim is to establish preliminary structure-activity relationships (SAR) to explore the physiochemical and ADMET properties of the compounds. ADMET refers what the body does to the drug in term of adsorption, distribution, metabolism and excretion. This phase lasts about 6 months.

The lead optimization is the most resource-intensive phase in drug discovery. The main challenge is to develop one or more compound with sufficient affinity for the target, acceptable

drug-like properties and efficient to work in the cell. Lead optimization takes 18-30 months. The output is a set of compounds with in vivo efficacy in animal models and with acceptable pharmacokinetics properties. Pharmacokinetics refers the study of how an organism affects a drug.

In preclinical trials the compounds are prepared in order to be tested in humans. This includes synthesis, formulation, toxicology and design of clinical trials. The synthesis and purification of compounds has a huge impact on the project cost.

Clinical trials are the most expensive and time-consuming process in drug discovery. They can be divided in three separate stages. In phase I the drug's safety is studied, in phase II drug's efficacy is tested and in phase III the number of patients is increased in order to evaluate drug's effectiveness, benefit and adverse reaction. If drug succeeds all phases successfully it is launched in the market. However, continued trials and monitoring is required[4,5].

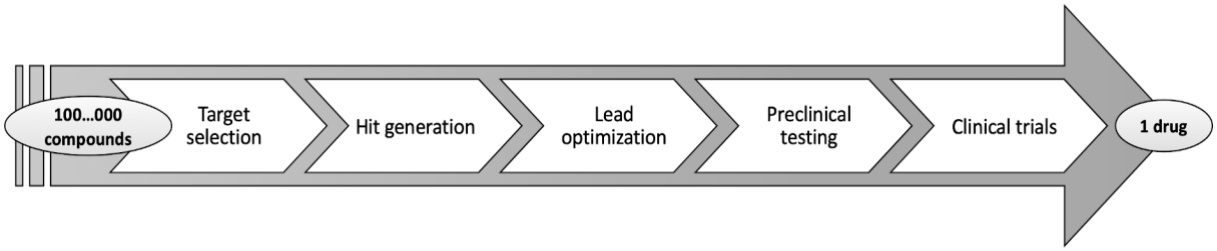


Figure 1. Drug discovery process. Protein-ligand affinity is investigated with computational methods in hit generation and with experimental methods in lead optimization in order to produce drug candidates for preclinical and clinical trials from initial library of millions of compounds.

2.2 Computer Aided Drug Design

Many pharmaceutical companies employ Hight Throughput Screening (HTS) in hit genration phase. HTS can identify molecules with chemical novelty with no a priori knowledge of the drug binding site on the target protein. Compounds are screened in cell-based assays to test for a change in the activity of specific signaling pathways. However, HTS is very expensive, consuming large quantities of target and compounds and requiring significant investment in robotic screening device. On the other hand, employment of computer-aided drug discovery (CADD) techniques by pharmaceutical companies became essential for the preliminary stage of drug discovery to expedite the drug development process in a more cost-efficient way and to minimize failures in the final stage. CADD approaches allows to reduce biological tests, to reject

compound with poor quality, to supply drug-receptor interaction pattern, to let a faster and more cost-efficient lead discovery and to provide compounds with high success rates[6,7].

In particular, the computational drug design tools can be divided into ligand-based drug design (LBDD) and structure-based drug design (SBDD) (Figure 2). These two methods are dependent on the information available on the identified target. SBDD approach uses 3D structure of the target for the generation or screening of potential ligands followed by synthesis, biological testing, and optimization. In contrast, LBBD approach employs computational modeling methods to develop theoretical predictive models of molecules with diverse structures and known potency. In detail, SBDD design and evaluate ligands based on their predicted interactions with the protein binding site[8]. In SBDD 3D proteins structure can be downloaded from an online dataset, as for example the Protein Data Bank website[2] (PDB). The protein structures usually are obtained with methods such as X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR). However, not all the protein structures that the human genome can encode are disclosed. In this case, homology modeling[9] permits to build new protein model by using templates (structures that are phylogenetically similar to the target). SBDD can be divided into two categories: de novo design and virtual screening. De novo design approach designs specific ligands for particular target based on the composition of the active site and the orientation of various amino acids at the binding site. Virtual screening approach uses available small molecules libraries to identify compounds with specific bioactivity for target biomolecules[10]. The aim of virtual screening is to rank active and inactive molecules. The commonly used strategy for virtual screening is molecular docking. Molecular docking models ligand-protein interaction in order to predict the ligand conformation in the active site (pose) and estimates the binding affinity.

In case where 3D protein structure is lacking, LBDD method is employed. The information extract from a set of ligands active against a target can be used to identify significant structural and physicochemical properties (molecular descriptors) responsible for the observed biological activity. Quantitative structure activity relationship (QSAR) and pharmacophore-based methods are common technique employed in LBDD. QSAR can be used to derive a model that correlates molecular structures with features responsible of biological activities. In order to build the model, the number of compounds with activity to consider should be greater than 20 and they should be acquired using the same experimental protocol. Moreover, molecular descriptors should have no autocorrelation to avoid overfitting and the final model should to be validate.

Alternatively, a pharmacophore model can be generated. The model should be contain information of molecules that bind to the biological target of interest[11].

Both SBDD and LBDD are successfully utilized in drug discovery. Another solution is the integration of these two methods in a drug discovery study. It can provide better and more extensive information in the modeling of innovative drug candidates against various diseases[12].

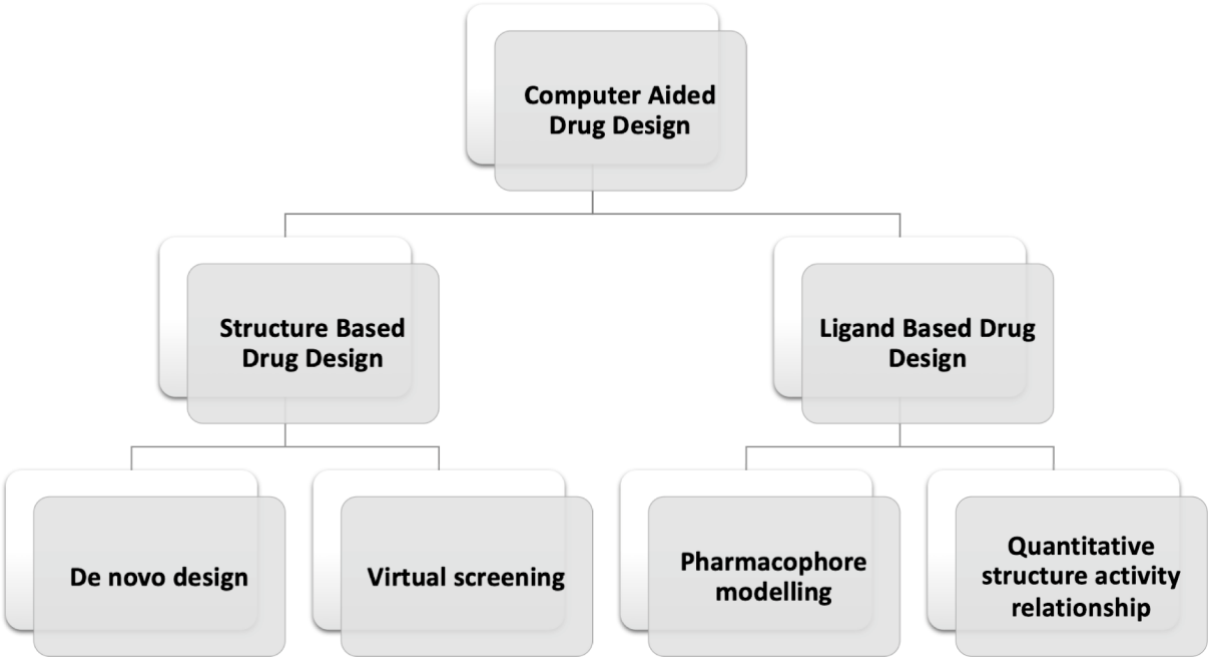


Figure 2. Workflow of Computer Aided Drug Design (CADD). There are two major types of drug design. The first is referred to as ligand-based drug design (LBDD) and the second as structure-based drug design (SBDB). In LBDD the ligand is known, and the structure of the protein remains unknown. The ligand is used to derive a pharmacophore model that defines the minimum necessary structural characteristics that a molecule must possess in order to bind to the target. Alternatively, the quantitative structure-activity relationship (QSAR) can be used. It correlates chemical structures with biological activity in a dataset of chemicals. In SBDB ligand and target are both known. SBDD can be divided in de novo design and virtual screening. In first method ligands are specified designed for the target. In the second method a compounds library is used in order to identify drug candidates.

2.3 Thermodynamics of Protein-Ligand Complexes

Proteins research represent a primary interest in biomedical science. Proteins involve in structural, immune, transport and enzymatic functions. In order to exert this function, they have to bind other molecules such as peptides, nucleic acids and ligands. The knowledge of the interaction binding is needed to understand biochemical process[13]. In particular, proteins-ligand binding is investigated in drug design in order to develop new molecules with pharmacological activity.

A drug works only when bound to its target receptor. The strength of the binding interaction between a biomolecule and a ligand is known as binding affinity. Quantify binding affinity allows to estimate the drug in vivo efficacy[14].

The region of the protein responsible of interaction with another molecule is known as binding pocket. Three different models have been proposed to explain the protein–ligand binding mechanisms. In Figure 3 the models of protein-ligand binding are displayed.

The first model called "lock-key" model, was introduced by Fisher[15] in 1894. In this model the protein and the ligand are considered rigid. The binding pocket is thought as a lock in which only the correct ligand size (the key) can be insert. Multiple conformations are neglected. For this reason, it may lead to wrong evaluations[16]. The second model is the induced fit model which considers the conformational flexibility of the ligand binding site[17]. The conformational select model is the third model that takes into account dynamic interactions. Protein does not exist as a single, rigid conformation but rather as an ensemble of conformational states that coexist in equilibrium. The ligand can bind selectively to the most suitable conformational state.

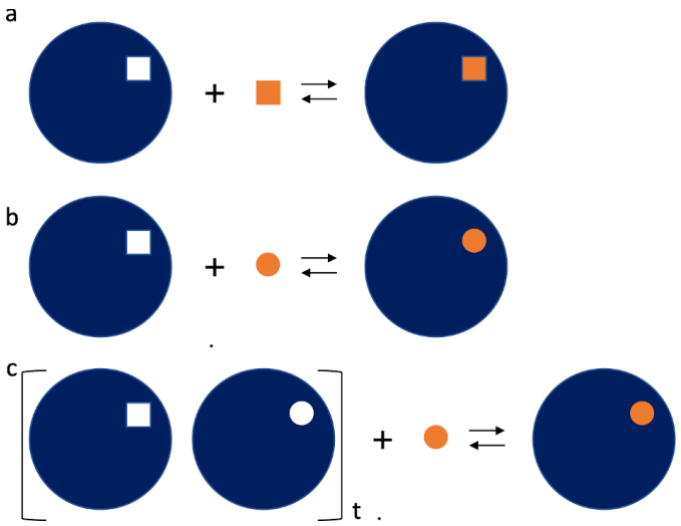


Figure 3. The three different binding models of protein and ligand. (a) Lock and key model, (b) the induced fit model and (c) the conformational selection model. The protein is illustrated in blue and the ligand in orange.

The ligand L and the protein P in the unbound states generate the complex LP after a chemical reaction as reported below:

$$P + L \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} PL \qquad [1]$$

where PL represents the protein-ligand complex, k_{on} and k_{off} are the kinetic rate constants that explain the binding and unbinding reactions, respectively. The units of k_{on} and k_{off} are $M^{-1}\,s^{-1}$ and s^{-1}, respectively.

Binding affinity is measured by the equilibrium binding constant K_b and dissociation constant K_d. In this context, smaller K_d values correspond to a greater binding affinity of the ligand for its target[18].

In simple terms, K_d corresponds to the drug concentration at which half of the receptor binding sites are occupied and is defined by the following concentration ratio of reactants and product:

$$K_b = \frac{k_{on}}{k_{off}} = \frac{[P][L]}{[PL]} = \frac{1}{K_d} \qquad [2]$$

the square brackets indicate the equilibrium concentration of protein [P], ligand [L] and protein-ligand complex [PL].

Therefore, the fast binding rate accompanied by a slow dissociation rate will give a high binding constant and, hence, a high binding affinity[16].

The binding affinity is influenced by non-covalent intermolecular interactions such as hydrogen bonding, electrostatic interactions, hydrophobic and Van der Waals forces between the two molecules.

Other important parameters useful to define affinity are IC50 or EC50, namely, the drug concentrations giving half-maximal inhibition or effect[13].

From an energetically point of view, binding affinity is calculated from the Gibbs free energy difference between the bound and unbound states. A helpful analytical way to calculate the binding free energy ΔG using the constant K_b or K_d is:

$$\Delta G = -RTlnK_b = RTlnK_d \qquad [3]$$

where R is the gas constant and T is the temperature at which the binding occurs. ΔG is a state function, it is completely determined by the initial and final state of the system, it does not rely from the path that connects the two states[19].

In Figure 4 the free energy difference between the bound and the unbound states is shown.

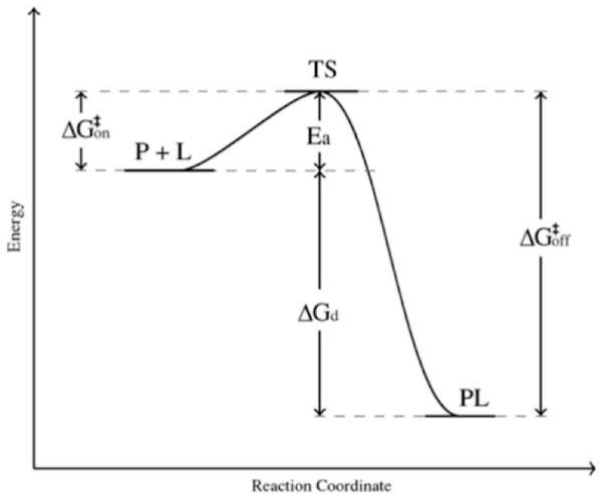


Figure 4. Free energy difference between the bound and the unbound states. P refers proteins, L ligand and PL the complex. TS is the transition state, E_a is the activation energy of the process, ΔG_d is the difference between the free energy of the reactants and of the product. ΔG_{on} is the free energy difference between the reactants and TS. ΔG_{off} is the free energy difference between the product and the TS[13].

Alternatively, the Gibbs free energy ΔG can be definite as:

$$\Delta G = \Delta H - T\Delta S \qquad [4]$$

where ΔH is the difference in enthalpy and ΔS is the difference in entropy of the system in the bound and unbound state.

Enthalpy is a measure of the total energy of a thermodynamic system. It is the sum of the internal energies of the solute and solvent plus the product of its pressure and volume. In exothermic process ΔH is negative while in endothermic process ΔH is positive. Exothermic process refers the formations of the energetically favorable noncovalent interactions between atoms and the endothermic process refers the disruptions of the energetically favorable noncovalent interactions. For a binding process ΔH reflects the energy change of the system when the ligand binds to the protein. The enthalpy considers the changes in Van Der Waals, hydrogen, electrostatic, polar, and aromatic interaction.

Entropy is a measure of how the heat energy is distributed over the thermodynamic system. The second law of thermodynamics determines that the heat always flows spontaneously from regions of higher temperature to regions of lower temperature. This reduces the degree of the order of the initial system, and, therefore, entropy could also be viewed as a measure of the disorder or randomness in atoms and molecules in a system. ΔS is a global thermodynamic property of a system, with its positive and negative signs indicating the overall increase and decrease in degree of the freedom of the system, respectively.

There are a lot of experimental techniques available to determine binding affinity of a ligand protein complex that can be divided by categories: stability shift assay, mobility shift assay, spectroscopic assay and calorimetric techniques[20].

Stability shift assay

These methods evaluate the ligand's effect on protein stability. Stability is assessed by denaturing the protein using temperature, pressure or chemical agents. The change in stability can be related to K_b and therefore ΔG. The method is used only for single-domain and single-binding site systems since the change of stability occurs independently in different domains, for this reason different signals produced can be difficult to interpret. An example is fluorescence thermal shift assay (FTSA)[21] in which protein is denatured increasing temperature at different ligand concentrations. The denaturation is evaluated by measuring the fluorescence of a dye molecule, which reports the regions of proteins unfolded.

Mobility shift assay

In this category the ligand is physically separate from the protein. The binding affinity is determined comparing bound fraction with target concentration. Molecule can be separated using force like centrifugal force in analytical ultracentrifugation method, electric field in electrophoresis and infrared laser-induced temperature gradients in microscale thermophoresis. Mobility shift assay are easy to use but the limit is that they cannot employed when the difference between free and bound and molecules is too irrelevant[22].

Spectroscopy assays

Spectroscopy assays connect the change in spectroscopic properties when ligand bounds protein with binding affinity. Same spectroscopic methods are fluorescence, dynamic light scattering (DLS)[23] and surface plasmon resonance (SPR)[24]. Fluorescence spectroscopy relates variation in fluorescence intensity with molecule binding. DLS detect the size of free and bound molecules in solution. Then, affinity is evaluated according with the ratio computed. In SPR affinity is obtained from the ratio between the binding rates k_{on} and the dissociation rates k_{off}. To calculate the constants, the protein is immobilized on the sensor surface and the ligand is free in solution, or conversely. The surface refractive index of the sensor changes when ligand bound the protein. The alteration is proportionally to the mass bound.

Finally, also the nuclear magnetic resonance (NMR)[25] is utilized to measure K_d. In NMR a magnetic field is applied to the compounds in the bound and unbound states. The magnetic field

affects the spins of the nuclei that release energy. The difference of energy between the two state is correlated to K_d.

Calorimetric techniques

Moreover, there are also calorimetric techniques such as Isothermal Titration Calorimetry (ITC) in order to study binding motive forces. ITC is the only approach to directly measure heat exchange during complex formation at constant temperature and has become the reference standard in determining the forces that guide the bonding process or stabilize intermolecular interactions. During an ITC experiment the ligand is inserted into a solution containing the protein of interest, the heat released or absorbed during their binding is measured in order to obtain the binding constant K_b.

Therefore, there is a great variety of experimental methods to evaluate the ligand-protein binding affinity and the recommended techniques depend on the considered system.

Binding free energy is an important feature of every receptor-ligand system useful to extract fruitful information on the complexation strength of the considered molecular system.

In this framework, computational methodologies to calculate the binding free energy can prefer faster but less precise or slower but more accurate approaches. Firsts one, are referred to the docking algorithms, while seconds one, are based molecular dynamics (MD) and Monte Carlo (MC) simulations[26].

Figure 5 shows the quality/speed ratio of the free energy calculating methods using computational techniques.

The category based on MD simulation can be divided into End Point Methods, Alchemical Modifications Method and Pathway Free Energy Methods. End Point Methods only require the bound and free states of the ligand. They include linear interaction energies (LIE)[27], molecular mechanics Poisson−Boltzmann surface area (MM-PBSA)[28], and molecular mechanics Generalized Born surface area (MM-GBSA)[29]. Alchemical Modifications Methods increase the accuracy of prediction by enhancing the sampling, including the free energy perturbation (FEP)[30] and thermodynamic integration (TI)[31] methods. Pathway Free Energy Methods, such as Umbrella Sampling (US)[32] and Steered MD[33], reproduce the dissociation path starting from the bound protein–ligand complex until the ligand physically separated from the protein receptor[34].

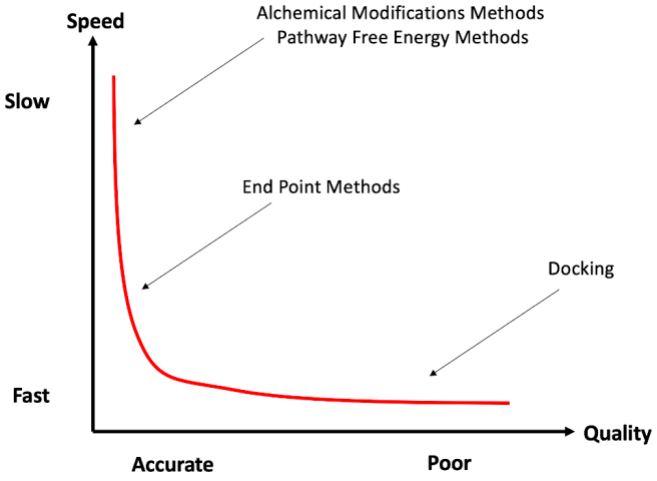


Figure 5: Methods to predict binding free energy can prefer speed or quality. Alchemical modifications and pathway free energy methods employ more computational time, but the result is more accurate than docking method which is faster but lower in quality. End point methods are a compromise between speed and quality.

LIE calculates the absolute binding free energy without sampling any intermediate state between the initial and final states, offering a good compromise between speed and accuracy. It considers the electrostatic and the van der Waals interaction between the ligand and the solvent for ligand in solution e for ligand in protein binding site[35].

$$\Delta G = \Delta G^{ele} + \Delta G^{vdw} = \text{a}(\langle E_{ele}^{L-S}\rangle_{PL} - \langle E_{ele}^{L-S}\rangle_{L}) + \text{b}(\langle E_{vdw}^{L-S}\rangle_{PL} - \langle E_{vdw}^{L-S}\rangle_{L}) \quad [5]$$

$\langle E_{ele}^{L-S}\rangle_{PL}$ and $\langle E_{vdw}^{L-S}\rangle_{PL}$ are ensemble averages of the electrostatic and van der Waals interaction energies between the ligand and the solvated protein from a molecular dynamics trajectory with ligand bound to protein. $\langle E_{ele}^{L-S}\rangle_{L}$ and $\langle E_{vdw}^{L-S}\rangle_{L}$ are ensemble averages of the electrostatic and van der Waals interaction energies of ligand in water. α and β are two empirical parameters that account for the internal energy of the solvent and the protein.

However, the method does not consider the energy of protein solvation and conformational entropy change of the ligand. Therefore, the best results are obtained when calculating the binding free energies for ligands with similar structures.

MM/PBSA computes binding free energy in trajectories generated from MD simulation for free ligand, free protein and their complex. Three energetic terms taken into account are: a) the potential energy in the vacuum, that includes bonded terms such as bond, angle, and torsion energies and nonbonded terms such as van der Waals and electrostatic interactions; b) the polar and nonpolar solvation energy; c) the configurational entropy.

$$\Delta G = G_{PL} - (G_P + G_L) \quad [6]$$

$$G = \langle E_{MM}\rangle + \langle G_{solvation}\rangle - TS \quad [7]$$

$\langle E_{MM} \rangle$ is the average standard molecular mechanics potential energy, $\langle G_{solvation} \rangle$ is the polar and non-polar contributions to the solvation free energies and the last term is the absolute temperature T multiplied by the entropy, S. It is computed for protein (P), ligand (L) and protein-ligand complex (PL).

Electrostatics and van der Waals interactions are modeled using respectively a Coulomb and Lennard-Jones (LJ) potential function. The free energy of solvation is the energy required to transfer a solute from vacuum into the solvent. The polar contribution is obtained by solving the Poisson-Boltzmann (PB) equation, whereas the non-polar term is estimated from a linear relation to the solvent accessible surface area (SASA) or solvent accessible volume (SAV). Entropy is estimated by a normal-mode analysis of the vibrational frequencies.

It represents a middle ground between the fast but very inaccurate docking and the accurate but time expensive FEP.

A major problem with MM-PBSA is the poor precision. Such a poor precision makes the method useless when comparing ligands with similar affinities or when comparing results obtained with different approaches or by different groups[36]. MM-PBSA is used to compare the binding free energy between different ligands towards the same target.

MM/GBSA, despite the MM/PBSA, uses the generalized Born (GB) model to compute the polar contribution to the free energy. The GB is an approximation which speeds up the treatment of the Poisson-Boltzmann equation[37]. Many studies have proved that GB calculation is much faster than the PB calculation but gives a less accurate result[38].

FEP computes relative binding free energy between two or more ligands toward one protein employing a thermodynamic cycle. In the thermodynamic cycle is displayed in Figure 6. There is ligand A and ligand B in the water and ligand A and ligand B bounded to the same protein. The relative binding free energy is calculated considering the difference between the free energy of transforming ligand A to ligand B in the protein and the free energy of transforming ligand A to ligand B in the solvent.

$$\Delta G = \Delta G_{bind}^{B} - \Delta G_{bind}^{A} = \Delta G_{1}^{A \to B} - \Delta G_{2}^{A \to B} \quad [8]$$

The advantage of FEP is high accuracy (in the order of 1 kcal/mol). On the other hand, if the binding mode is not known, is not possible to apply this method. Furthermore, it can be used only if there are very structurally similar ligands[26].

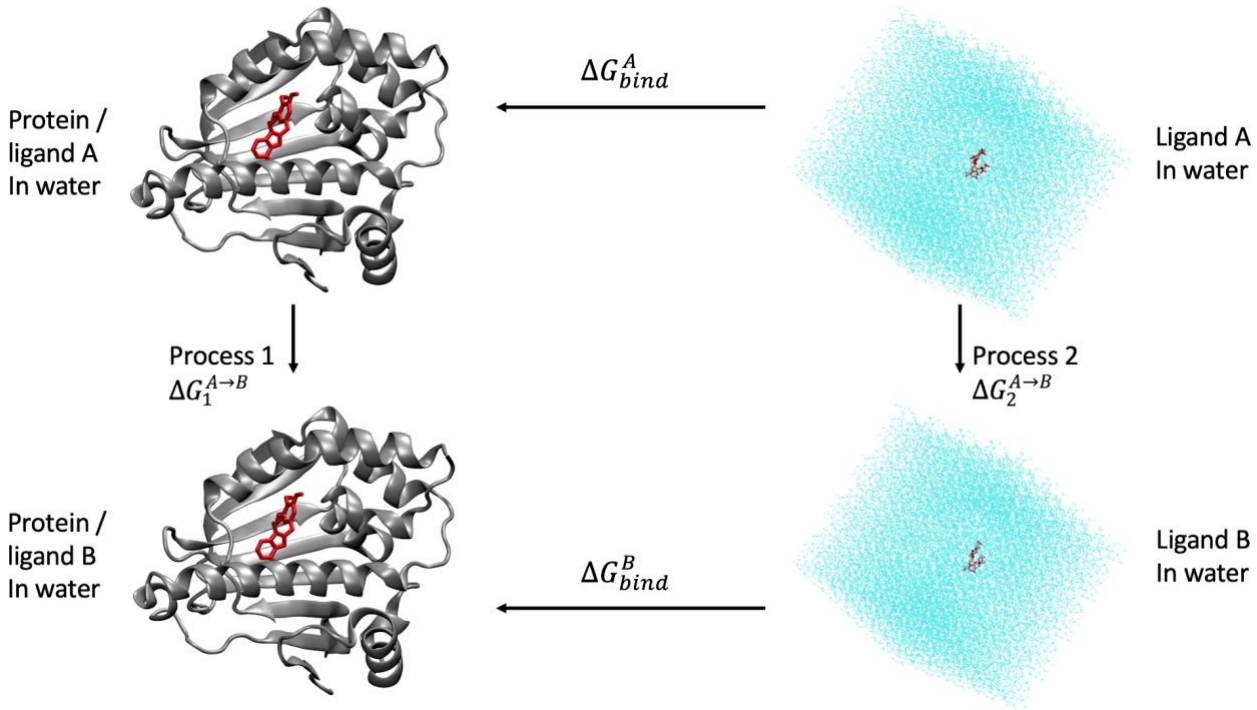


Figure 6. The thermodynamic cycle used to calculate the relative binding free energy in FEP method. In Process 1 the free energy of transforming ligand A to ligand B in the protein in estimated; in process 2 the free energy of transforming ligand A to ligand B in the solvent in obtained. The relative binding free energy is the difference between the free energies computed from process 1 and 2 and it is related to the relative binding free energy difference between the two ligands. Protein is shown as gray cartoon, ligands are represented as red licorice and the water solvent as cyan surface.

TI applies the alchemical transformation to turn from the bound state into the unbound using a coupling parameter λ that varies from 0 to 1. 0 refers the initial state and 1 the final state.
The final free energy is given by the sum of all the transformations between different states.

$$\Delta G = \int_0^1 \langle \frac{\partial H(\lambda)}{\partial \lambda} \rangle \, d\lambda \qquad [9]$$

TI is accurate but computationally challenging because it requires an extensive sampling of intermediate states[39].

US computes binding free energy from the potential mean force (PMF). An external bias potential is applied to the system to drive the ligand from the bound to unboned state. The system is forced to sample regions of conformational space that would not otherwise be accessible. Subsequently, the pulling trajectory is divided in a series of windows, which cover the entire pathway and a bias restrained MD calculation is performed in each window. The result is a series of histograms, which contain the biased distribution of the reaction coordinate from each window. There histograms are then unbiased and combined usually with the aid of

the weighted histogram analysis method (WHAM). The PMF is calculated by the WHAM equations[40].

Steered MD is a MD simulation in which is applied a force to the ligand in order pull it out from the binding site of the protein. The PMF and the binding free energy are computed with the Jarzynski nonequilibrium work theorem[41]. In practice, it is possible to compute the difference binding free energy computing the work done. The averaged work in pulling trajectory is related to the difference binding free energy[42].

3 Materials and Methods

This chapter provides a theoretical overview of the methods employed in the present work. Concept of molecular modelling for investigating biological mechanisms is introduced. Then, molecular docking is discussed, focusing on search algorithm and energy scoring function for generating and evaluating ligand poses. Finally, Molecular Mechanics approach is introduced to provide a background on physical basis behind molecular modelling.

3.1 Introduction to Molecular Modelling

Molecular modelling includes all theoretical methods and computational techniques used to model and mimic the complex behavior of molecular systems. Molecular systems are composed by a huge number of molecules therefore is not trivial to analytically evaluate the properties of the system. In order to address this issue, systems can be studied using numerical methods. Thanks to the increasing power of computers, it is possible to simulate biological systems with millions of atoms in a reasonable amount of time.

The most detailed analysis of a biological system is at the quantum level, where each electron-electron interaction is considered. This accurate representation is obtained by solving the Schrödinger equation. However, it is not feasible solve Schrödinger equation for any system containing more than one hundred atoms because the complex nature of the interactions and for the long times it would require. For this reason, a simplified molecular description of the system is taken into consideration. The system is defined in terms of interaction between atoms and it is governed by Newton's laws. The forces depend to the deformation of chemical bonds, hydrogen bonding, electrostatics, and van der Waals interactions.

In contrast to the quantum level, simulations of relatively large dynamic systems are possible at the molecular level in a reasonable time.

If the system contains a large number of atoms and is required to be simulated for a significant length of time, Coarse Grained (CG) method can be used. In CG, the number of degrees of freedom in the system is reduced by treating groups of particles as single entities, allowing longer simulations with the same computational effort, but with a reduction in the accuracy of the results.

Therefore, molecular modeling techniques are employed to understand biological systems that are often challenging to obtain with laboratory analysis, besides experiments in biological field are expensive in both money and time. For example, computational techniques play a valuable role in pharmaceutical research. In general, some common MD applications are in protein folding/unfolding, drug delivery, polymers chains analysis, transport and diffusion properties, protein free energies, polymer aggregation, multiscale modelling and much more.

3.2 Molecular Docking

Molecular docking is one of the most well-known SBDD methods employed in discovery and design of new drugs. The aim of molecular docking is to predict the experimental binding mode and affinity of a small molecule within the binding site of the receptor target of interest. A search algorithm and an energy scoring function are the basic tools of a docking methodology for generating and evaluating the ligand conformations.

3.2.1 Ligand-Protein Docking

Molecular docking includes protein-protein or ligand-protein interactions[43]. Ligand-protein docking (Figure 7) is a computational method used to predict the best fit orientation of a ligand that binds to a receptor. Most docking programs can rank the activity of each compound by analyzing the different ligand-target interactions and estimating the binding affinity of the complex[44,45].

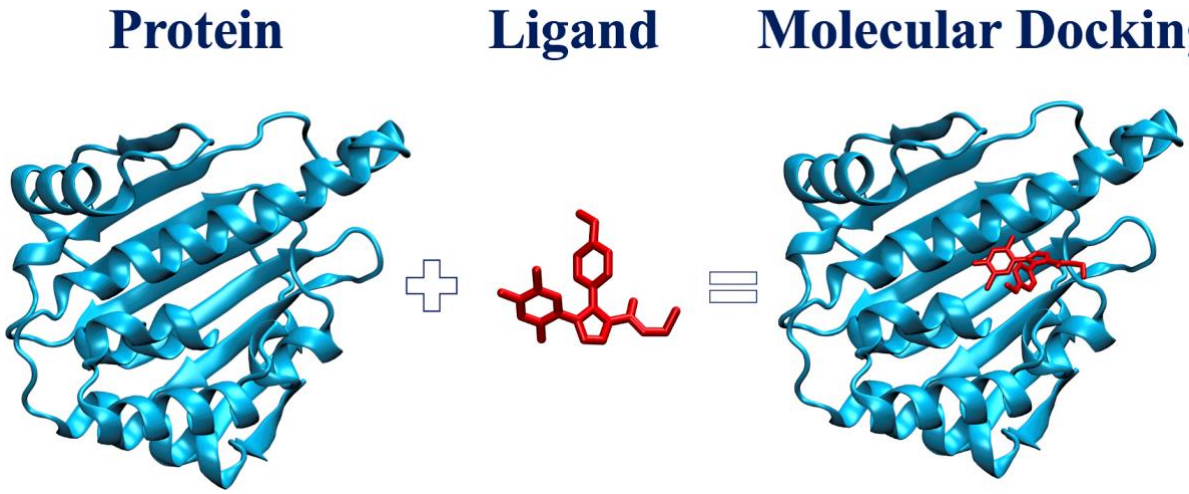


Figure 7. Elements involved in Molecular Docking: protein, ligand and the complex. Protein is represented in cyan and ligand in red.

In order to perform a docking procedure, the target structure can be download from Protein Data Bank website. Crystal structures with high resolution have to be chosen. It is suggested a resolution value less than 2Å[44].

Models of ligands are available on database such as ZINC15[46], PubChem[47], DrugBank[48]. If the structure are not present in the databases is possible design the small molecule using a design software such as Avogadro[49], ChemDraw[50]. Another possibility is to employ protein structure crystallized with the ligand already inserted in the correct pose.

The above-mentioned solution is useful to validate the performance of the docking algorithm since the correct position of the ligand is known.

Before executing docking procedure, the protein and the ligand have to be prepared. The basics steps consist of calculating the protonation states and assign atomic partial charges[51].

Furthermore, if the binding site is already known, it must be selected and delimited, otherwise the whole protein's surface will investigate, burning a lot of computation time.

Docking can be of two types: rigid docking and flexible docking. In particular it is possible to consider both ligand and protein rigid, flexible ligand and rigid protein or both ligand and protein flexible. The rigid docking is based on the lock-and-key assumption proposed by Fischer. Both the ligand and the receptor cannot change their spatial shape, new conformations are not generated. It is allowed only its rotation and translation. Protein-ligand affinity is directly proportional to a geometric fit between their shapes[52]. In the flexible docking bond angles, bond lengths and torsion angles of the components are modified. It is based on the induced-fit theory proposed by Koshland [53].

The success of docking algorithms is normally measured in terms of the root-mean-square deviation (RMSD) between the coordinates of experimentally ligand conformation and the predicted by the algorithm. A good performance is usually considered when the RMSD is less than 2Å[51].

In general, the aims of docking studies are to identify the ligand pose in the active site and to estimate the correct affinity value. Serval search algorithm are been developed to achieve the first purpose, and several scoring functions for the second (Figure 8).

3.2.2 Search Algorithm

The aim of the search is to identify ligand pose with the lowest energy. The search space consists in all possible binding modes between protein and ligand. However, it is impossible to explore the whole conformational space, but only a small amount of it can be sampled.

Three main search methods may be identified: systematic, stochastic or by using simulation methods as Molecular Dynamic (MD) and Monte Carlo (MC). In systematic search all rotatable ligand bonds are gradually rotated in order to cover all possible combinations among the dihedral angles. Systematic algorithms can be divided into two classes: exhaustive search and fragmentation algorithms. The first method ideally rotates all possible ligand bonds to explore all the possible conformations. The second approach divides the ligand into several fragments that are separately docked in the receptor site. For example DOCK[54], FLOG[55] and LUDI[56] software uses the fragmentation algorithms. Using systematic method is more likely that the algorithm converges to the local minimum rather than the global minimum. To overcome this issue, it is preferable to start from different conformations, rather than use the same starting position.

Stochastic exploration samples the conformational space of a ligand by generating random variations in the orientation of all rotatable bonds and in some cases random translations for the whole ligand within the binding site. The algorithm is more suitable for large molecules, where the degrees of freedom are too high for an efficient systematic search. It is more probable to find a global minimum, but on the other hand the computational cost increases[45].

Some common algorithms that apply stochastic approach are Monte Carlo (MC), Simulated annealing (SA)[57] and Genetic Algorithms (GA)[58]. MC method generates ligand-protein complexes by performing random changes in the ligand conformation. Each obtained complex is energetically analyzed, and the most favorable state is selected. Unfortunately, it tends to be trapped in local minimum, therefore simulated annealing has been applied to avoid this issue. MCDOCK[59] is a program that utilizes MC method.

In SA the temperature is varied during the run for each conformation in order to better explore the conformational state. AutoDock[60] uses simulated annealing approach for more accurate conformation exploration. GA performs the search taking a cue from evolutionary processes. Different translations, orientations and conformations of the ligand are encoded as binary strings called genes. These genes compose the 'chromosome' which represents the pose of the ligand. An initial population of chromosomes are generated in order to cover a wide area of the

energy landscape. Genetic operators, like mutations and crossovers, are applied to the population to obtain new ligand structure. New structures will be assessed by scoring function, and the ones that survived can be used for the next generation. This enables evolution of optimal solutions that represents the correct binding mode. Genetic algorithms have been successfully used in molecular docking programs such as Gold[61] and AutoDock[62].

Simulation methods like MD are more accurate but at the same time computationally demanding. Indeed, these methods are not the best choice to analyze a huge number of compounds. MD samples the conformational space in order to obtain statistically relevant macroscopic information from the ensemble. However, MD are often unable to cross high energy barriers, which results in sampling only the local minima derived from the starting conditions. Different strategies can overcome this issue, for example by using simulated annealing or Metadynamics[63]. DOCK perform a minimization step after each fragment addition.

3.2.3 Scoring Function

The scoring functions are mathematical methods that rank ligand poses in order to predict binding affinity, quantifying several ligand-protein interaction types such as hydrogen bond, electrostatics, van der Waals's forces and hydrophobic interactions. There are three main types of scoring functions: Force field-based methods, Empirical scoring functions and Knowledge-based scoring functions.

Force-field based method estimates the binding energy considering the contribution of bonded (bond stretching, angle bending and dihedral variation) and non-bonded terms (electrostatics and van der Waals interaction). It has some limitations because it does not consider solvation and entropic contributions. Moreover, it needs to include a cut-off distance to treat non-bonded interaction[64] which leads in a loss of accuracy.

Empirical scoring functions computes the binding energy summing energetic factors that concerned in ligand-receptor complex formation. They are hydrogen bonds, ionic interactions, hydrophobic and entropic effects, etc. The energetic factors are multiplied by a rescaling coefficient obtained from a linear regression analysis of a training set of complexes with known binding affinities. This method is faster than force-field-based method but its limitation is that the result strongly rely on the training set used [45].

Knowledge-based scoring functions generates a function considering pairwise potentials derived from known ligand-receptor complexes. These potentials are calculated considering the

frequency of two different atoms that are found within a given distance in the structural dataset. The final score is the sum of these interactions. The main advantage of using knowledge-based functions is the computational simplicity, which can be applied to easily screen huge compound libraries. However, some limitations are inherent in the limited training sets of protein–ligand complex structures[65].

Consensus scoring is a procedure that combines information from different approach in order to reduce limitation of each method.

Lately, a new approach has been developed to improve the performance of the above-mentioned scoring function. This is the Machine-learning-based scoring functions which are used for rescoring purposes, in order to enhance the accuracy. Machine-learning-based scoring functions employ machine-learning algorithms, such as support vector machine, random forest, neural network, deep-learning, etc[64].

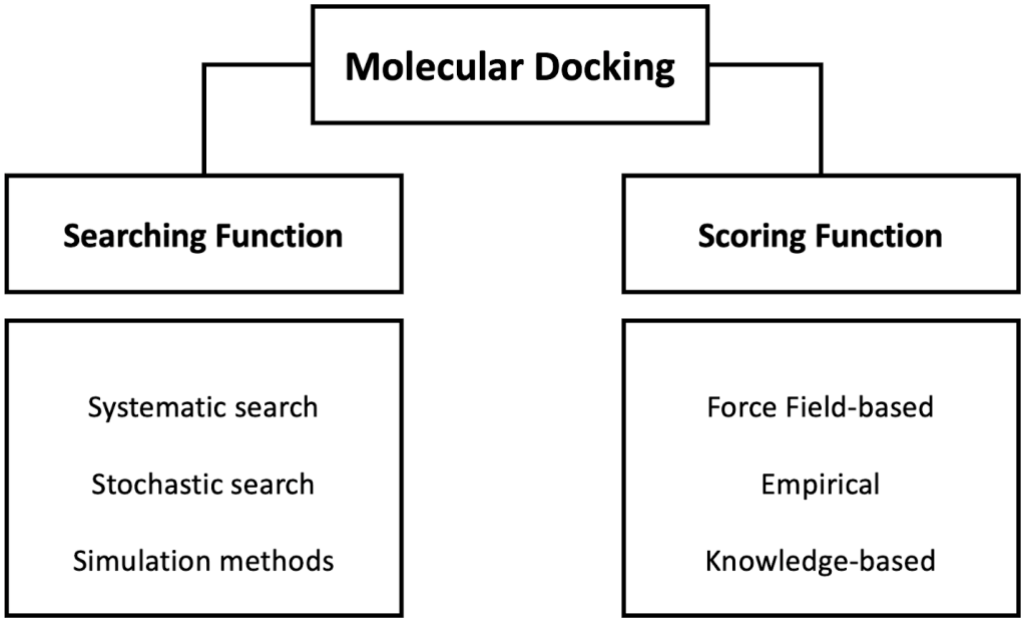


Figure 8: Docking protocol. It can be described as a combination of a search algorithm and a scoring function. Searching functions can be divided in systematic search stochastic search and search by using simulation methods. Scoring function are force field based, empirical and consensus.

3.2.4 Docking Software

Today, there are at least 60 docking programs commercially (or freely) available with different force fields, conformational sampling algorithms and a variety of scoring functions[52].

The most commonly used programs are AutoDock[60], GOLD[61], Glide[66], DOCK[54], AutoDock Vina[67] ICM[68] and FlexX[69]. A more exhaustive list is shown in Table 1 and in

Table 2. The subdivision has been made according to the conformational search method and the implemented scoring function.

Table 1: Docking software listed according to the conformational search method.

Systematic search	Stochastic search
SLIDE[70]	PLANTS[71]
DOCK[54]	EADock[72]
FlexX[69]	MOE_Dock[73]
FRED[74]	Gold[61]
GLIDE[66]	ICM[68]
EUDOC[75]	LigandFit[76]
Surflex-Dock[77]	PRO_LEADS[78]
Hammerhead[79]	CDocker[80]
Flog[55]	GlamDock[81]
ADAM[82]	MolDock[83]
eHiTS[84]	AutoDock[60]

Table 2: Docking software listed according to the implemented scoring function.

Force-Field-Based	Empirical	Knowledge-Based
Gold[61]	GLIDE[66]	PMF_Score[85]
DOCK[54]	LUDI[86]	MotifScore[87]
AutoDock[60]	ChemScore[88]	PoseScore[89]
LigandFit[76]	LigScore[90]	DrugScore[91]
MedusaScore[92]	HYDE[93]	SMoG[94]
ICM[68]	PLP[95]	PESD_SVM[96]

AutoDock4 is docking software which uses a Lamarckian genetic algorithm (LGA) to generate multiple poses of a ligand in a pocket. AutoDock4 docking performs many independent LGA runs

followed by clustering based on the root mean square deviation (RMSD) of the resulting poses to identify the most populated portion of the conformational space of the ligand. The AutoDock4 scoring function is an empirical scoring function which computes the non-bonded interaction potential including van der Waals contribution, hydrogen bon term, electrostatic interaction and desolvation potential. Each term is multiplied by an empirical coefficient obtained from a calibration of a training dataset of bound complexes with known binding affinities. The total score of a binding pose is obtained by adding the difference of intra-molecular energies between the protein-ligand bound and unbound forms, then subtracting the difference of inter-molecular energies. Finally, a simple entropic term is introduced to consider the variation of the entropy of the system. The success rate of AutoDock4 is around 53% in reproducing crystallographic poses.

Another popular docking software is DOCK. DOCK uses fragment-based algorithm to generate ligand poses in a binding site, where the ligand is placed in the pocket and the flexible branches are sequentially grown around them. DOCK applies a force-field based scoring function which models non-bonded interactions between ligand and protein atoms as a sum of Lennard-Jones 12-6 and electrostatic terms. The pose success rate is around 73%.

Glide is a docking software based on an exhaustive systematic search algorithm used to sample the ligand conformational space, followed by a minimization step. The scoring method is a combination of a force-field-based function, an empirical function and the strain energy of the ligand conformation. The success rate of Glide is around 66%.

A well-known tool for protein-ligand docking is AutoDock Vina (Vina). Vina uses an iterated local search global optimizer searching method to generate poses of the ligand within the binding site. The scoring function employed by Vina combines aspects from knowledge-based and empirical potentials considering steric, hydrophobic, hydrogen bonding and entropic terms.

The scoring function is:

$$c = \sum_{i<j} f_{t_i t_j}(r_{ij}) \qquad [10]$$

where the summation is over all atoms separated by three consecutive covalent bonds. A type t_i and a set of interaction function $f_{t_i t_j}$ of the interatomic distance r_{ij} is assigned to each atom. The value is given by a sum of intermolecular and intramolecular contributions. Empirical information from both the conformational preferences of the receptor-ligand complexes and the experimental affinity measurements are extracted. The values of experimental affinity

measurements are derived from the PDBbind[97] dataset. The optimization of Vina algorithm consists of uses not only the value of the scoring function but also its gradient. The gradient is obtained by deriving the scoring function with respect to the position, orientation, and torsions of active rotatable bonds of ligand. Vina outperforms AutoDock4 in both accuracy and speed. Indeed, it is able to identify the correct binding pose in 78% of the cases[98,99].

3.3 Molecular Mechanics

The term Molecular Mechanics refers the application of classical mechanic to determinations of molecular equilibrium structures. The structures studied vary from small molecules to large biological system.

Molecular systems are modelled by Newtonian mechanics with a series of approximations in order to reduce the complexity of the system. In this context, according to the Born-Oppenheimer approximation[100], atoms are treated as spheres and bonds are considered as springs, neglecting the electronic motions.

3.3.1 Potential Energy Function

The potential energy of the systems is estimated using a set of equations and parameters known as force field (FF). The total potential energy is the sum of the potential energy of binding and the potential energy of non-binding interaction:

$$V = V_{bonded} + V_{non-bonded} \quad [11]$$

The bonded interactions consider the variation of bond lengths, angles and dihedrals. Instead, the non-bonded interactions are given by the van der Waals potential and Coulomb electrostatic potential.

$$V_{bonded} = V_{bonds} + V_{angles} + V_{dihedrals} \quad [12]$$

$$V_{non-bonded} = V_{electrostatic} + V_{van\ der\ Waals} \quad [13]$$

3.3.2 Treatment of Bond and Non-Bond Interactions

The potential energy function for a molecular system composed of N atoms identified by a vector position r_i can be described as (Figure 9):

$$V(r_1, r_2, \dots, r_N)$$

$$= \sum_{bonds} \frac{1}{2} k_l (l - l_0)^2 + \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} k_\varphi (1 + \cos(n\varphi - \delta))$$

$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right) + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}] \qquad [14]$$

The first term refers the covalent bond between two atoms. Bonds are modelled as a harmonic interaction, where k is the force constant, l_0 is the reference bond length (the length when all the other FF terms are zero) and l is the bond length (length at the equilibrium minimum when all the terms are considered). The second term indicates the interactions between three atoms, modelled as a harmonic interaction. k_θ is the force constant, θ_0 is the reference bond angle (the angle assumed when all the other FF term are zero) and θ is the bond angle (the angle at the equilibrium minimum when all the terms are considered).

The third term is for dihedral angles which originates between four atoms. The energy related to the dihedral angle is modelled as a series of cosines, where k_φ is the energy cost related to the dihedral angle deformation, n is the number of energetic minima along a complete rotation and δ is the minimum position for the torsional angle. Dihedrals can be divided into proper and improper dihedrals. First refers when full rotation is allowed, second when the rotation is limited.

The last term defines the non-bonded interaction, modelled as functions inversely proportional to the distance between two atoms. This term includes Van der Waals forces and Coulomb electrostatic interactions.

Van der Waals potential is the weakest intermolecular force and occur among atoms with no net electrostatic charge. Van der Waals forces are modelled with Lennard-Jones (L-J) equation:

$$V_{L-J} = 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \qquad [15]$$

σ is the collision diameter, the minimum distance with the interaction potential equal to zero, and ε is the well depth, the interaction potential energy minimum. The term raised to twelfth refers the interactions that act at long range as attractive force. The component raised to the sixth describes the short-range interaction that play a role as repulsive force in order to avoiding the overlap between atom.

Electrostatics interaction develop among pairs of non-bonds charged atoms. This force is described by Coulomb's law:

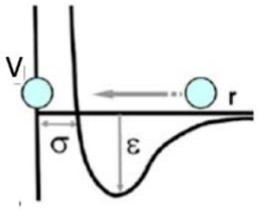
$$V_e = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \qquad [16]$$

ε_0 is the vacuum electrical permittivity, r_{ij} is the distance between the charge q_i of the atom i and the charge q_j of the atom j.

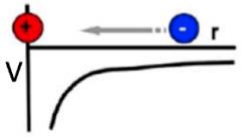
This interaction is definite as long-range interaction because the energy decreases as the distance between two atoms is reduced.

Calculation of non-bonded interactions requires an expansive computational effort. To reduce this problem, the non-bond interactions are computed by applying a cutoff distance. It allows to compute the interaction only if the distance between atoms is smaller than the cutoff.

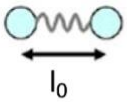
$$V = \sum_{i=1}^{N}\sum_{j=i+1}^{N} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$



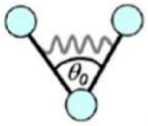
$$+ \sum_{i=1}^{N}\sum_{j=i+1}^{N} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$



$$+ \sum_{bonds} \frac{1}{2} k_l (l - l_0)^2 +$$



$$+ \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_0)^2 +$$



$$+ \sum_{dihedrals} k_\varphi [1 + \cos(n\varphi - \delta)]$$

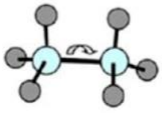


Figure 9. Potential energy function for molecular interactions. The first term represents the non-bond interaction of Van der Waals force, the second the non- bond interaction of Coulomb force and the last three terms describe the bond interactions of bond, angles and dihedral respectively.

3.3.3 Periodic Boundary Conditions

All the system atoms are contained in a three-dimensional box filled with implicitly or explicitly model of water. Several possible boxes are cubic, parallelepiped, hexagonal prism, octahedron and dodecahedron. Box boundaries are a crucial issue for MD simulations which may strongly affect the properties of the whole system. The implementation of periodic boundary conditions (PBC) allow avoiding side effect. The system is surrounded by a number of identical boxes. In

this manner, the particles near the boundary of the main box can feel the interactions with the particles in the next periodic box. The number of particles in the main box remains constant because if a molecule leaves the box during the simulation it is replaced by its own periodic image that comes into the box from a neighbor (Figure 10).

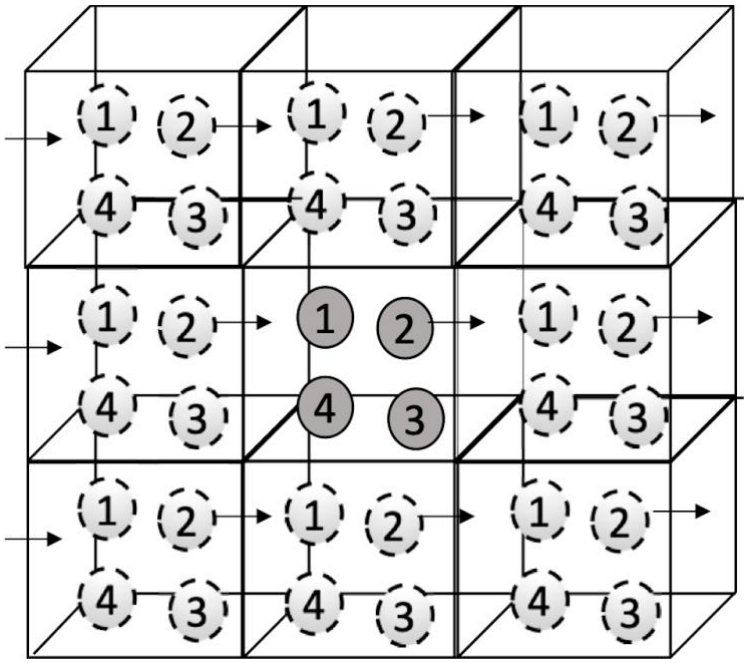


Figure 10. Periodic boundary conditions. The first box is in center and then it is surrounded by copy of itself.

3.3.4 Potential Energy Minimization

The Potential Energy surface (PES) is a complicated multidimensional function of the molecular system coordinates. PES is characterized by stationary and saddles points. Firsts are local or even global minima and refer the more stable states of the system. The second are the highest point between minima and belong to the transition state. The aim of energy minimization is to find an arrangement of atoms that corresponds to the local minimum energy of the system.

There are several algorithms that are able to perform an energy minimization in order to identify the minimum point of the PES: the derivative and the non-derivative methods.

An example of non-derivative method is the simplex[101] algorithm. On the other hand, derivative methods can be divided into first-order methods such as Steepest Descent and Conjugate Gradient and second-order methods such as Newton-Raphson[102] and L-BFGS. In general, first-order approaches use the direction of the first derivate of the energy (the gradient) to indicate where the minimum lies. Whereas second-order approaches use both first and second derivatives. The second derivative gives information on the curvature of the PES hence

predicting where the function will change direction. These methods have a main common limitation: they can only go downhill on the PES, thus they will find the closest minimum from the starting point that could be a local minimum.

4 Improving Accuracy of Blind Protein-Ligand Docking by Fragmented Docking Method

This chapter presents a novel protocol used to perform protein-ligand docking. Protein-ligand docking is the procedure able to predict the position and orientation of a ligand when it is bound to a receptor. Furthermore, the top discovered hits have been reported with a detailed discussion about the validity of the approach.

4.1 Introduction

Discovery and development of a new drug is generally known as a very complex process which requires a lot of time and resources. The main goal of drug discovery is to obtain compounds that powerfully interact with therapeutic targets[10]. Computer Aided Drug Design (CADD) approaches are widely used to increase the efficiency of the drug design. CADD techniques are essential for the preliminary stage of drug discovery to expedite the drug development process in a more cost-efficient way and to minimize failures in the final stage. In particular, computational drug design tools can be divided into ligand-based drug design (LBDD) and structure-based drug design (SBDD). The choice between these two methods dependents on the identified biological target information available. SBDD approach uses 3D structure of the target for the generation or screening of potential ligands. In contrast, LBBD approach is employed where 3D protein structure is lacking and it utilizes computational modeling methods to develop theoretical predictive models of molecules with diverse structures and known potency[8].

In this framework, molecular docking is one of the most powerful techniques of SBDD. The aim of molecular docking is to predict the experimental binding mode and affinity of a small molecule within the binding site of the selected target receptor. A search algorithm and an energy scoring function are the basic tools of a docking methodology for generating and evaluating the ligand conformations. Most applications of docking are performed by knowing the protein binding region. In this case, a small docking box is selected around the protein binding site in order to facilitate the docking by focusing sampling of the translational, rotational, and torsional degrees of freedom of the ligand. However, there are situation in which the information about binding site is missing and it becomes necessary to explore the entire protein surface by docking algorithms. Several methods have been developed to overcome the problem

of not recognizing the binding site. AutoDock Vina (Vina)[67] is a docking software efficient when the binding site is known. When the protein binding pocket is unidentified, Vina can execute the so-called blind docking (BD). In BD, the target is included into a single research box and the correct pose of the ligand is sought on the entire surface of the protein. This method has many limitations because it is improbable to exhaustively sample the whole energy landscape in a fixed number of steps to find the best ligand conformation[103]. Another approach to address the above-mentioned issue consists in reducing the search space, focusing only in some areas of the protein. The method employs the SiteHound algorithm to predict the location of potential binding site and after that, it carries out multiple independent docking procedures on smaller boxes centered on predicted binding sites[104]. The results show that the docking focused on a small number of predicted binding sites reduces the computational time required to obtain the solution and generates results more accurate in terms of correct pose prediction in comparison to BD. However, if the real binding site is not included in the boxes in which the docking is carried out, the procedure could lead to an incorrect result. In overall, the main problems highlighted by the two analyzed methods consist in a too inaccurate sampling box for the BD method and uncertain prediction of the binding pocket for the SiteHound algorithm. In this work, a fragmented docking method (FD) has been developed to improve the performance of the previously discussed methods. The idea is to slice the docking box into multiple smaller boxes and then to merge all the results. Ligand-protein docking is carried out separately in each box with Vina, shifting the location of the box step by step, in order to cover the entire surface of the protein. The partition in several boxes allows the systematically exploration of the whole protein surface, which improves the discovering of ligand conformations adopted within each examined box. In addition, the complete investigation of the whole protein structure, intrinsically leads to discover of the real binding site.

The molecular docking calculations has been performed whit FD method on 116 crystal structures of Heat Shock Protein 90 – alpha (Hsp90α) and 176 of Human Immunodeficiency virus protease 1 (HIV-1 PR).

The developed method shows better performances than the standard methods employed to overcome the problem of not recognizing the binding site.

4.2 Material and Methods

4.2.1 Proteins Dataset

The two proteins employed in docking are Heat Shock Protein 90 – alpha (Hsp90α) and Human Immunodeficiency virus protease 1 (HIV-1 PR).

Heat Shock Protein is a chaperone protein that assists other proteins to fold correctly and allows cells to survive in extremely heat conditions. It is able to preserve the integrity of the cells if they are exposed to high temperatures by regulating the flux of calcium ions, maintaining the chromosomal stability and safeguarding the endoplasmic reticulum proteins homeostasis. In particular, Hsp90 indicates Hsp protein with a weight of roughly 90 kilodaltons. If on the one hand, Hsp90 aid the human body to survive at elevate temperature, on the other hand, is able to stabilize the proteins necessary for tumor growth. For this reason, Hsp90 inhibitors are mainly investigated as anti-cancer drugs[105]. Hsp90α is an isoform of Hsp and it is located in the cytoplasm of the eukaryotic cells. Hsp90α is a globular protein and contains secondary structures as alpha helixes, beta sheets and random coils. It consists of four structural domains: N-terminal domain, middle region, C-terminal domain and a linker region that connect the first two domains. The binding pocket is situated in the N-terminal domain, which is predominantly constituted by hydrophobic residue. The pocket is a binding site for many molecules included antibiotics like radicicol and geldanamycin, which have anti-tumor activity. Furthermore, it shows a high affinity for ATP[106,107].

HIV-1 PR is a retroviral aspartyl protease, an enzyme which acts in peptide bond hydrolysis in retroviruses, which is necessary for the life cycle of the retrovirus HIV that causes AIDS.

HIV-1 PR is constituted from 99 amino acid and exists as a homodimer with only one active site. The binding pocket is located between the identical subunits. Each monomer consist of a wide β-sheet region (a loop of glycine) which partly constitutes the binding site of the substrate and one of the two fundamental residues of aspartyl, Asp-25 and Asp-25' which are at the bottom of the cavity[108,109]. The HIV-1 PR enzyme activity can be inhibited from HIV protease inhibitors by blocking the active site of the protease.

4.2.2 Dataset Preparation

FD and BD methods have been carried out on the same set of complexes extracted from the Protein Data Bank (PDB)[2]. The dataset of Hsp90α is composed of 116 crystal structures, while the dataset of HIV-1 PR is composed of 176 crystal structures. As normal procedure, all the waters molecules have been removed from each PDB entry. Missing residues and atoms have been added to the protein chain utilizing the program Modeller[110]. The particular cases where the protein or ligand have a PDB record of double spatial positions of the atoms have been treated selecting only one of the positions pair. Hydrogens and Gasteiger charges have been added both the ligand and the protein employing python script prepare_ligand4.py and prepare_receptor4.py from MGLTools[60,111].

4.2.3 Docking procedure

Docking procedure has been performed using the software Vina with FD and BD method. Ligand and protein have been separated and prepared as described earlier for both approaches. In BD protocol a single docking experiment has been carried out on the whole protein surface, whereas in FD protocol multiple smaller docking experiment has been performed. The choice of the boxes number depends on the size of the protein and the respective ligand. In detail, the dimension of each box has been chosen twice the maximum ligand length and the overlap between different boxes has been set to fifty percent. Several docking operations have been performed in each box, which has been shifted step by step along the surface of the protein. Ten different ligand conformations have been generated from each docking process.

Results of FD and BD methods have been compared by root mean squared deviation (RMSD) of ligand heavy atoms of the solution for FD and DB with respect to the experimental ligand structures. In literature a value of RMSD lower or equal to 0.2 nm is recommended as a limit value for a good pose reproduction[112,113,114,115,116].

The RMSD comparison has been made employing the ligand configuration to which Vina attributed as the best in the affinity value ranking.

4.2.4 MM/GBSA rescoring

Docking programs generate binding poses of compounds in the active site of a target and evaluate the protein-ligand binding affinity by means of scoring functions. However, docking

scores and experimental binding affinities usually do not correlate, because screening large numbers of compounds in a reasonable time requires the use of approximate scoring function. Hence, docking results can be improved employing post-docking processing strategies. The MM/GBSA methods based on binding free energy estimations has been utilized to rescoring the docking results. MM/GBSA calculations were performed using MMPBSA.py[117] program, selecting the pairwise GB model developed by Hawkins et al. (GB^{HCT}; igb = 1 in Amber's terminology)[118]. The calculation has been performed in ten energy minimized ligand-protein complexes. Structures refer proteins bounded with the top scored docking ligand poses. Predicted binding free energies were compared with docking affinity results. To obtain a successful rescoring, it is necessary that the complex which the MM/GBSA attributes as the best due to the highest affinity value is the same in which the RMSD between the generate ligand and the original one is less than 0.2 nm.

4.3 Results

In this section are presented the results regarding the FD and BD docking procedures, comparing the performance of both methods. As pointed out before and illustrated in Figure 11, the main idea behind the FD method is to divide the exploration of the protein surface into smaller independent docking boxes. The advantage that result from using the slice and shift method to identify the original binding mode of a ligand will be shown below.

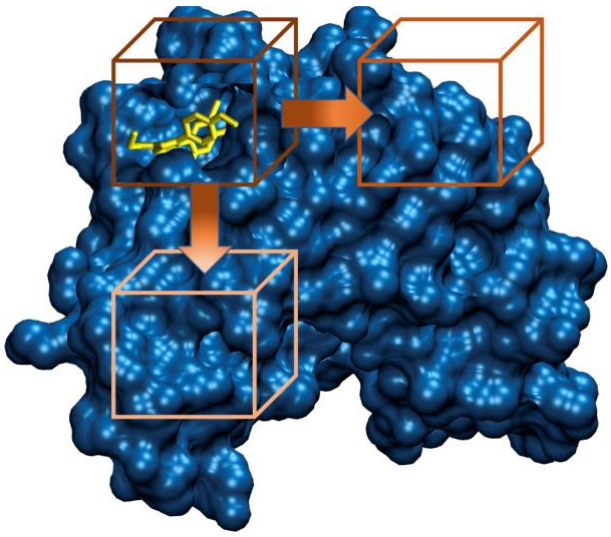


Figure 11. Main idea of FD: a docking procedure is performed in each box which moves on the whole protein surface. Protein is represented in blue, ligand in yellow.

4.3.1 Fragmented Docking versus Blind Docking

The first step to compare BD and FD method is to determine whether the docking results identified the correct binding mode of the ligands in the crystal structures. The poses with the lowest docking energy have been selected and the RMSD between the heavy atoms of docked ligand and the heavy atoms of the ligand in the crystal structure has been calculated. This measures the ability of the docking protocol to identify the correct ligand binding site in the target structure. In addition, it is also helpful to verify if the docking methods select the correct ligand conformation pose, once the binding site location is found. Successful docking is achieved if the RMSD between the docking pose and the ligand in the crystal structure will be lesser or equal to 0.2 nm.

The tables with the calculated RMSD values using FD and BD method of 116 Hsp90α and 176 HIV-1 PR proteins with co-crystallized ligands are reported in Supporting Information (Table S 1,Table S 2).

4.3.1.1 Protein-Ligand Complexes of Hsp90α

The first structure analysed is Hsp90α by performing a docking procedure on 116 different ligand-protein complexes. In Figure 12 are presented Hsp90α protein and 3 different ligands: radicicol, geldanamycin and ATP. In particular, the illustration aims to highlight the main interactions between the ligand and the active site of the protein. The above-mentioned ligands own a high affinity for the target protein. In this context, geldanamycin or radicicol are pharmacological inhibition of Hsp90α.

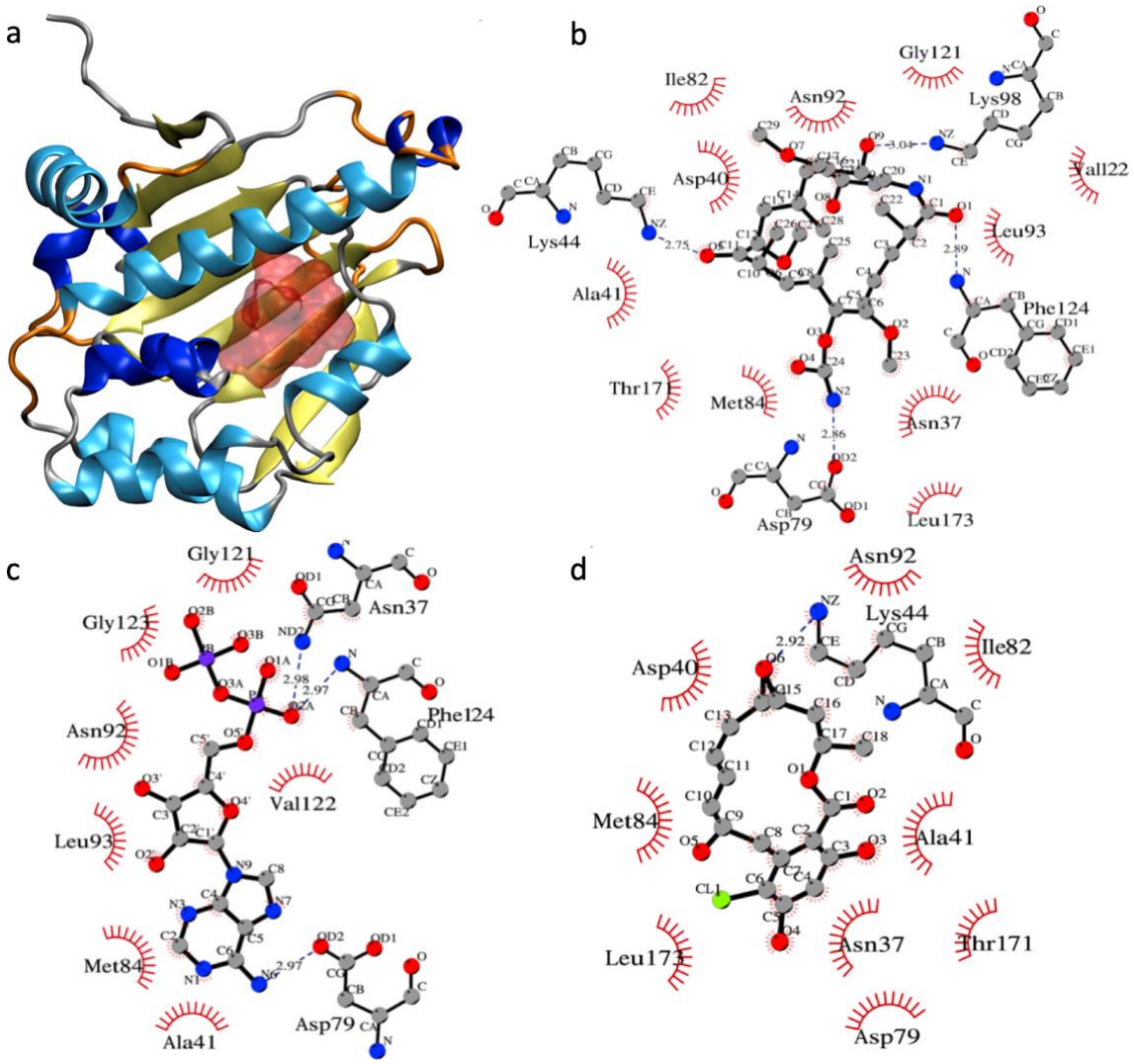


Figure 12. (a) NewCartoon structures of the Hsp90α with red surf of pocket binding site in N terminal domain. Alpha helixes are painted in cyan, 3_{10} helixes in blue, beta sheets in yellow, turns in orange, random coils in silver; (b) Ligplot diagram interactions between Hsp90α and geldanamycin (PDB: 1a4h); (c) Ligplot diagram interactions between Hsp90α and ADP (PDB: 1amw); (d) Ligplot diagram interactions between Hsp90α and radicicol (PDB: 1bgq). Dashes represent hydrogen bound between protein residue and ligand. The other protein residues mentioned are responsible for hydrophobic interactions.

As described in the previous sections, ten different ligand poses have been generated from each docking procedure. Therefore, the total number of ligands configuration produced are ten multiplied the number of boxes. Taken together all the docked configuration, the pose considered for the validation is the one which Vina has attributed the maximum affinity value with the protein. Since RMSD is the parameter used for the validation, a value of 0.2 nm is selected as threshold that distinguishes the correctly predicted poses from the wrong ones. Figure 13 shows several examples of generated ligand poses overlapped with the ligands in the crystal structures. In detail, the figure represents three examples of successful docking and three

examples of failed docking. In particular, the RMSD value are 0.02 nm, 0.09 nm, 0.16 nm, 0.23 nm, 0.32 nm, and 0.51 nm, respectively.

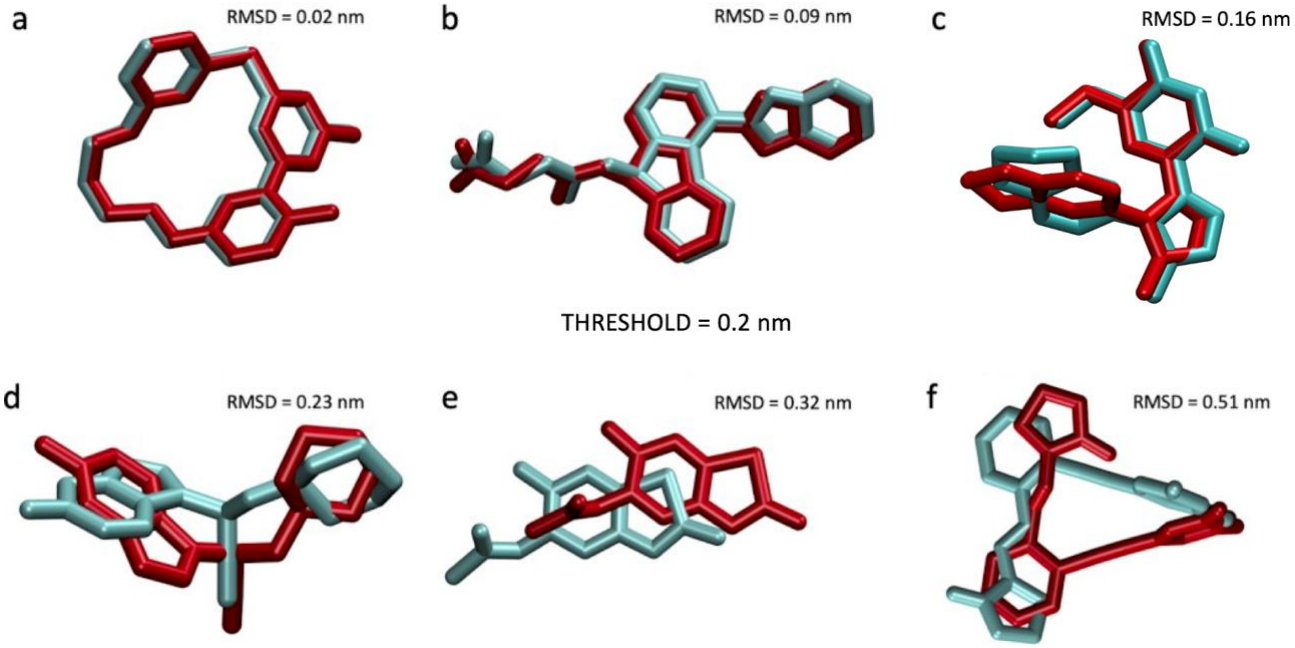


Figure 13. Comparison between generated by Vina and original ligand conformation. Original pose is in cyan and output pose is in red. (A) PDB: 3vha. RMSD computed is 0.02 nm; (B) PDB: 2ykb. RMSD computed is 0.09 nm; (c) PDB: 2brc. RMSD computed is 0.16 nm; (D) PDB: 4eft. RMSD computed is 0.23 nm; (E) PDB: 5fnd. RMSD computed is 0.32 nm; (F) PDB: 2qg2. RMSD computed is 0.51 nm.

The RMSD between the generated ligand configurations and the ligands in the crystal structures were calculated in order to compare the performance of FD and BD methods. In order to have a clear picture of the FD and BD performance, pie chart is selected as graphical tool to highlight the accuracy of both methods. The pie chart in Figure 14 shows the percentages of generated ligand poses with RMSD value lesser and greater than 0.2 nm in comparison to the original ligand conformation. Considering the output configuration with greatest affinity, BD finds the correct pose prediction in only 26.7% of attempts, while FD finds the correct prediction in the 62.1% of attempts.

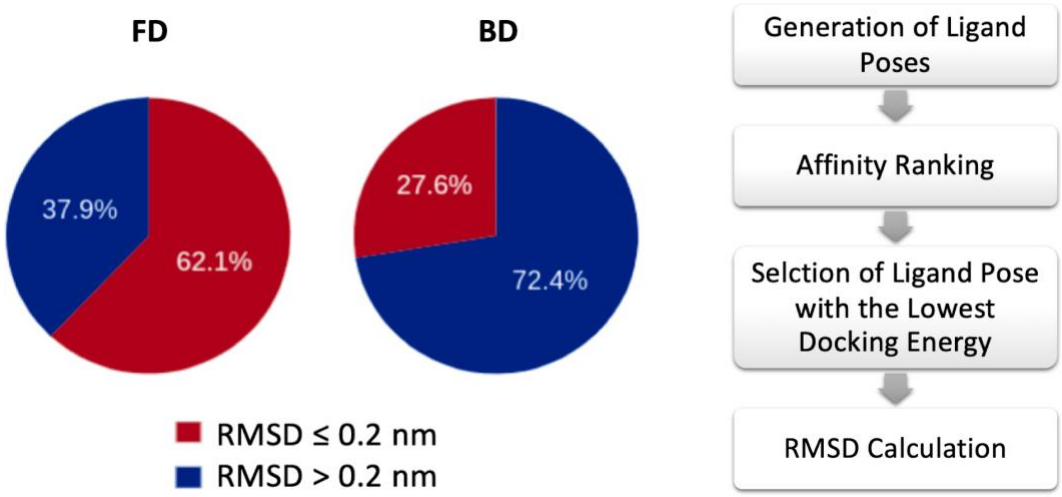


Figure 14. Percentage of poses generated with RMSD lesser and greater than 0.2 compared the original pose using the FD and BD method. The comparison has been made considering the poses with the maximum affinity value. Results of FD methods are shown in the pie chart on the left and results of BD method are shown in the pie chart on the right. Red indicates the correct configuration found and blue the wrong ones. An accurate ligand configuration has been discovered in 62.1% of cases using FD method and 27.6% using BD method.

From another point of view, Figure 15 shows a histogram of the distributions RMSD values. The size of bins is set to 0.2 nm and the frequency refers to the number of ligand conformation with that particular value of RMSD. As shown by the histogram, the interval between 0 and 0.2 nm contains the greatest number of outcomes for FD method: 72 are the correct poses predicted. While, the greatest number of outcomes for BD method are in the interval between 0.2 and 0.4 nm. The number of correct poses predicted with BD are 32. From these first results, it is possible consider that the FD method is able to produce ligand poses that are more accurate than those produced by BD.

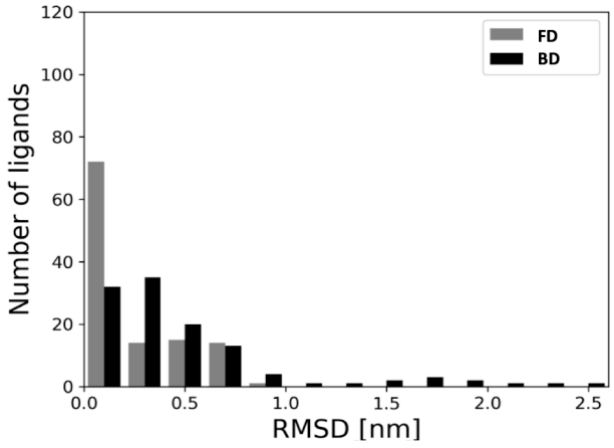


Figure 15. Histogram shows the distribution of the computed RMSD value between the output ligands with best affinity and the original. In FD method most of the generated configurations belong at the range between 0 and 0.2 nm. They are 72. In BD method most of the generated configurations belong at the range between 0.2 and 0.4 nm. They are 32.

The ligand conformation which Vina assigns the greatest affinity is not always the one that closest to the real pose of the ligand. In this framework, the first ten configurations have been selected and the RMSD analysis between them and the original ligand has been computed. The hypothesis is to find the pose with the lowest RMSD value between the first 10 records of Vina's output configurations. To complete the picture, Figure 16 shows the pie chart percentages of the obtained results considering the pose with lowest RMSD between the first 10 poses. Interestingly, an improvement in the performance is notable for both FD and BD methods. In detail, FD finds in 91.4% of times the poses with RMSD value lower than 0.2nm, while BD discovers in 43.1% of times the correct configuration.

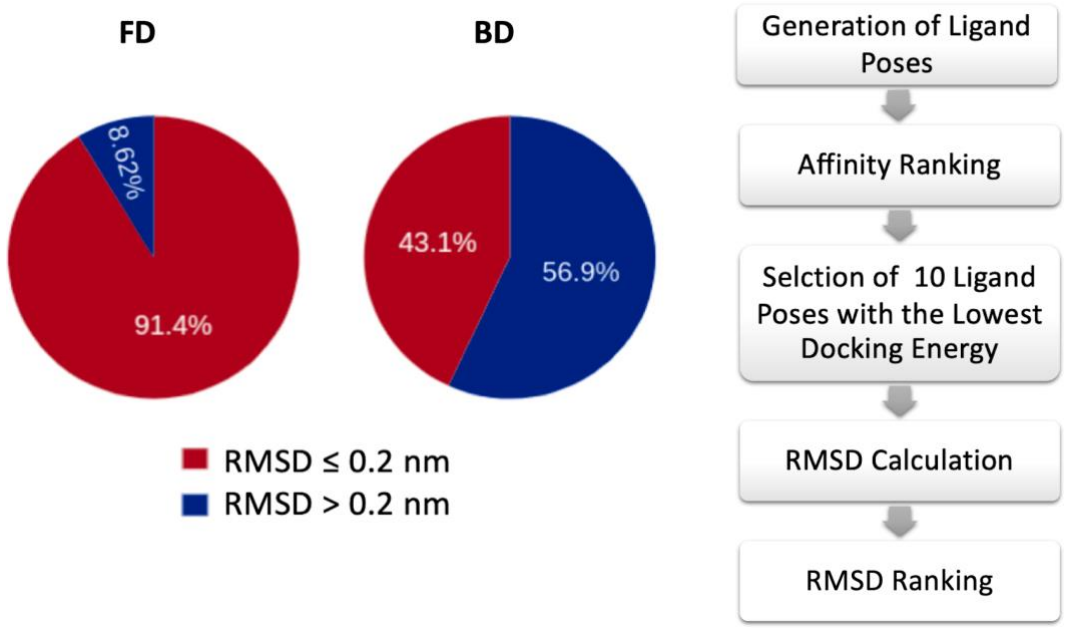


Figure 16. Percentage of poses generated with RMSD lesser and greater than 0.2 compared the original pose using the box and blind method. The comparison has been made considering the poses with the best RMSD value between first 10 with the maximum affinity value. Results of FD methods are shown in the pie chart on the left and results of BD method are shown in the pie chart on the right. Red indicates the correct configuration found and blue the wrong ones. An accurate ligand configuration has been discovered in 91.4% of cases using FD method and 43.1% using BD method.

As pointed out in the previously analysis, the distribution of RMSD values is shown in Figure 17. Compared to the previous case, the number of conformations belonging to the first interval has increased, indicating that is the correct poses have been predicted, as expected. The maximum frequency is 106 in FD and 50 in BD.

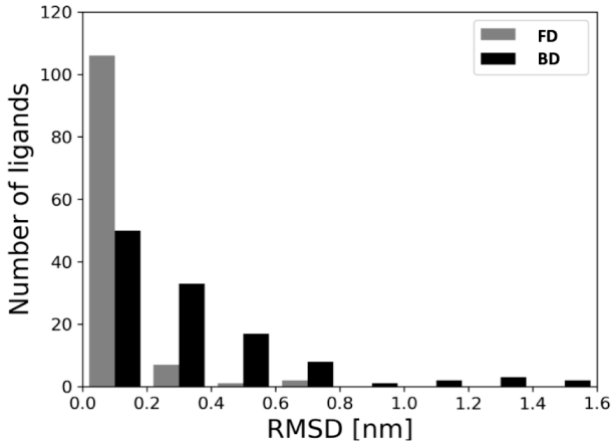


Figure 17. Histogram shows the distribution of the computed RMSD value considering the ten poses generated with greater affinity. In both method most of the generated configurations belong at the range between 0 and 0.2 nm. They are 106 in FD method and 50 in BD method.

In Figure 18 the relationship between the improving in performance with the increase in the number of poses considered is exhibited. The affinity ranking has been made including 1, 3, 5, 10, 20, 40, 70, 100 ligands configurations. The curve significantly rises from 1 to 10 poses considered, as expected. This suggests that Vina does not always attribute the highest affinity to the pose that closest to the real pose of the ligand. For this reason, a rescoring procedure has been performed in order to evaluate if MM/GBSA is able to attribute the best affinity value to the configuration with the lowest RMSD value. The results of aforementioned rescoring procedure will be discussed in 4.3.2 section.

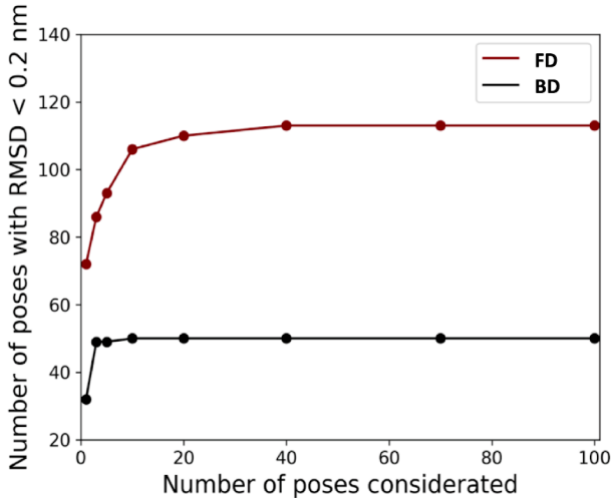


Figure 18. The curve underlines how the performances vary as the number of poses to consider calculating the RMSD increases. The greatest growth of the curve occurs when it goes from 0 to 10 in FD methods and from 0 to 5 in BD method. Red curve refers the FD method, black curve to BD method.

4.3.1.2 Protein-ligand complexes of HIV-1 PR

The second structure considered for the Vina FD and BD method is HIV-1 PR protein. In this context, the docking procedure has been performed 176 different ligand-protein complexes. Figure 19 shows HIV-1 PR protein and 3 different ligands: saquinavir, indinavir and ritonavir inhibitors. In particular, the illustration aims to highlight the main interactions between the ligand and the active site of the protein. The above-mentioned ligands are shown due to their high affinity for the target protein.

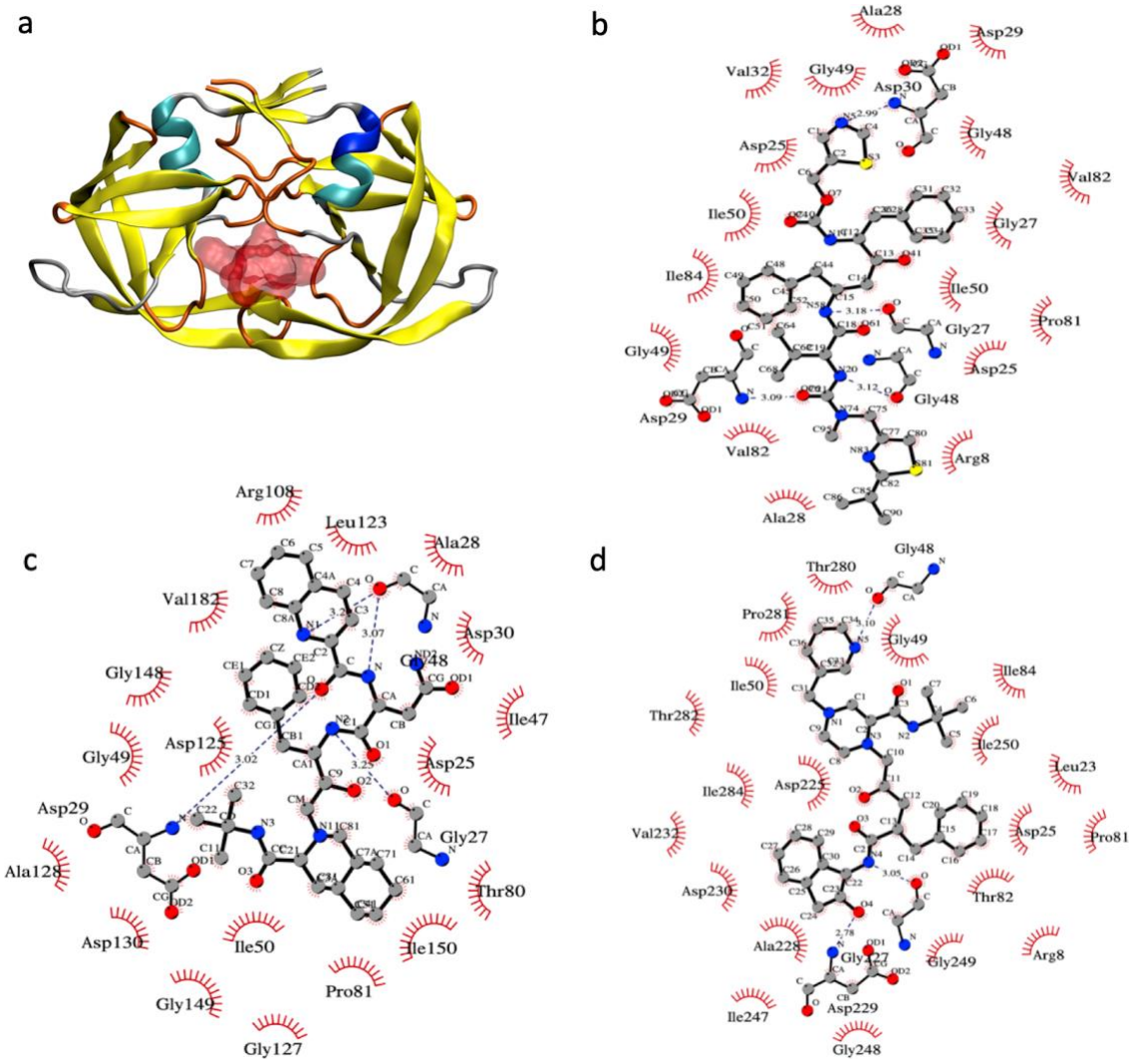


Figure 19. (a) NewCartoon structures of the HIV-1 PR with red surf of pocket binding site. Alpha helixes are painted in cyan, 3_{10} helixes in blue, beta sheets in yellow, turns in orange, random coils in silver; (b) Ligplot diagram interactions between HIV-1 PR and ritonavir (PDB: 1hxw); (c) Ligplot diagram interactions between HIV-1 PR and saquinavir (PDB: 2nnp); (d) Ligplot diagram interactions between HIV-1 PR and indinavir (PDB: 1c6y). Dashes represent hydrogen bound between protein residue and ligand. The other protein residues mentioned are responsible for hydrophobic interactions.

As applied before for the Hsp90α protein, ten different ligand poses have been generated from each docking procedure. Therefore, the total number of ligands configuration produced are ten multiplied the number of boxes. Taken together all the docked configuration, the pose considered for the validation is the one which Vina has attributed the maximum affinity value with the protein. Since RMSD is the parameter used for the validation, a value of 0.2 nm is selected as threshold that distinguishes the correctly predicted poses from the wrong ones. Figure 20 shows several examples of generated ligand poses overlapped with the ligands in the crystal structures. In detail, the figure represents three examples of successful docking and three examples of failed docking. In particular, the RMSD value are 0.03 nm, 0.1 nm, 0.19 nm, 0.22 nm, 0.46 nm and 1 nm, respectively.

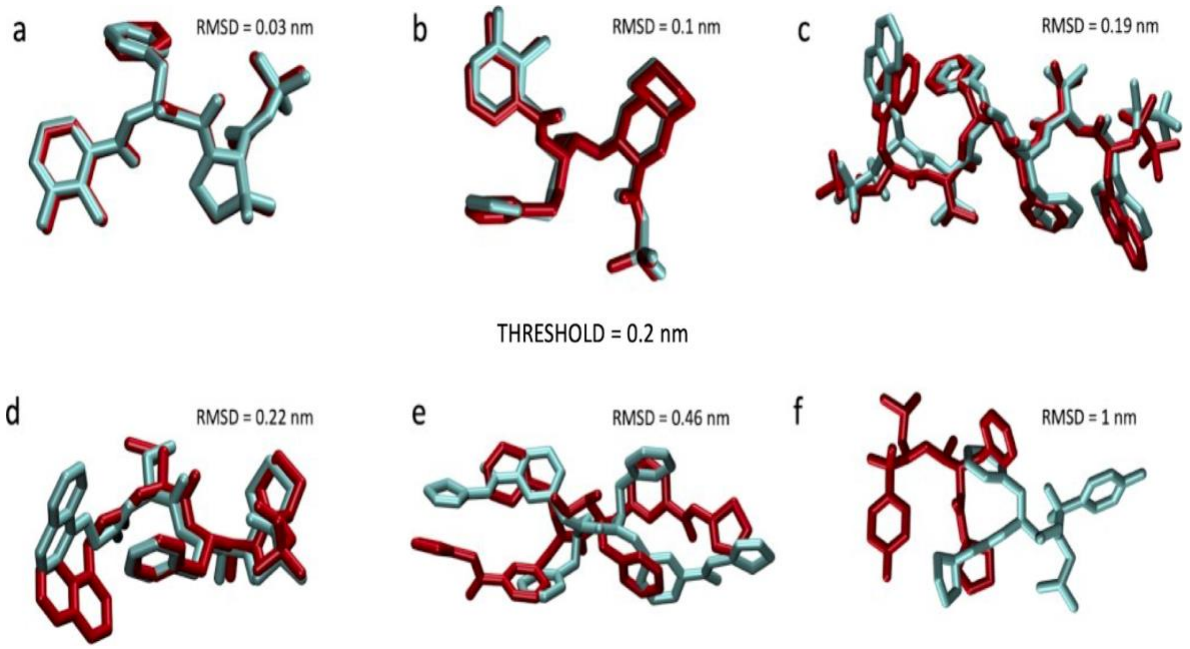


Figure 20. Comparison between generated by Vina and original ligand conformation. Original pose is in cyan and output pose is in red. (A) PDB: 1mrx. RMSD computed is 0.03 nm; (B) PDB: 3ekx. RMSD computed is 0.1 nm; (C) PDB: 1vik. RMSD computed is 0.19 nm; (D) PDB: 1hpx. RMSD computed is 0.22 nm; (E) PDB: 1bv9. RMSD computed is 0.46 nm; (F) PDB: 3ekv. RMSD computed is 1 nm.

The RMSD between the generated ligand configurations and the ligands in the crystal structures were calculated in order to compare the performance of FD and BD methods. The pie chart in Figure 21 shows the percentages of generated ligand poses with RMSD value lesser and greater than 0.2 nm in comparison to the original ligand conformation. Considering the output configuration with greatest affinity, BD finds the correct pose prediction in only 31.3% of attempts, while FD finds the correct prediction in the 64.8 % of attempts.

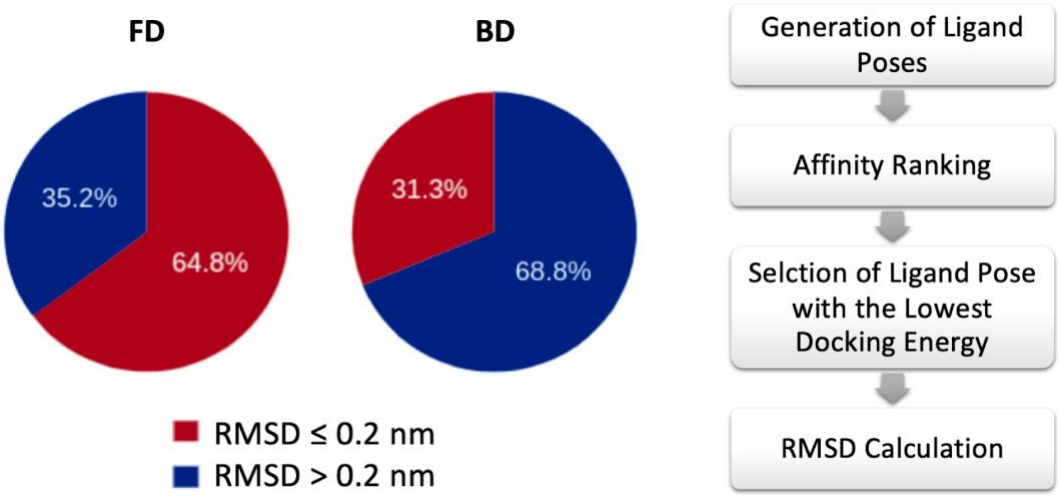


Figure 21. Percentage of poses generated with RMSD lesser and greater than 0.2 compared the original pose using the box and blind method. The comparison has been made considering the poses with the maximum affinity value. Results of FD methods are shown in the pie chart on the left and results of BD method are shown in the pie chart on the right. Red indicates the correct configuration found and blue the wrong ones. An accurate ligand configuration has been discovered in 64.8% of cases using FD method and 31.3% using BD method.

From another point of view, Figure 22 shows a histogram of the distributions RMSD values. The size of bins is set to 0.2 nm and the frequency refers to the number of ligand conformation with that particular value of RMSD. As shown by the histogram, the interval between 0 and 0.2 nm contains the greatest number of outcomes: 114 are the correct poses predicted with FD and 55 with BD. From these first results, it is possible consider that the FD method is able to produce ligand poses that are more accurate than those produced by BD.

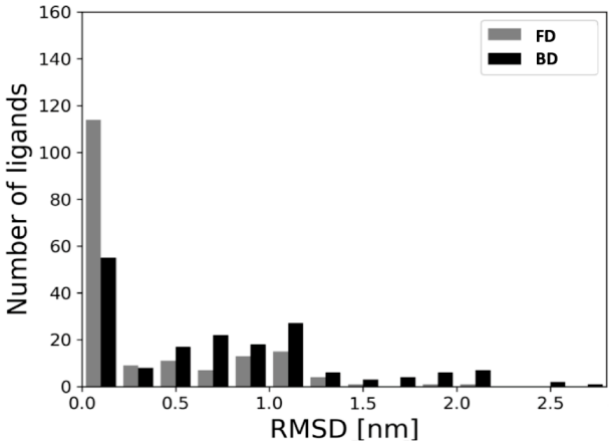


Figure 22. Histogram shows the distribution of the computed RMSD value between the output ligands with best affinity and the original. In FD method most of the generated configurations belong at the range between 0 and 0.2 nm. They are 114. In BD method most of the generated configurations belong at the same range between 0 and 0.2 nm. They are 55.

Considering also this second dataset, the ligand conformation which Vina assigns the greatest affinity is not always the one that closest to the real pose of the ligand. In this framework, the first ten configurations have been selected and the RMSD analysis between them and the original ligand has been computed. The hypothesis is to find the pose with the lowest RMSD value between the first 10 records of Vina's output configurations. Figure 23 shows the pie chart percentages of the obtained results considering the pose with lowest RMSD between the first 10 poses. There is an improvement in the performance for both box and blind methods. In detail, FD finds in 84.7% of times the poses with RMSD value lower than 0.2nm, while BD discovers in 41.5%. of times the correct configuration.

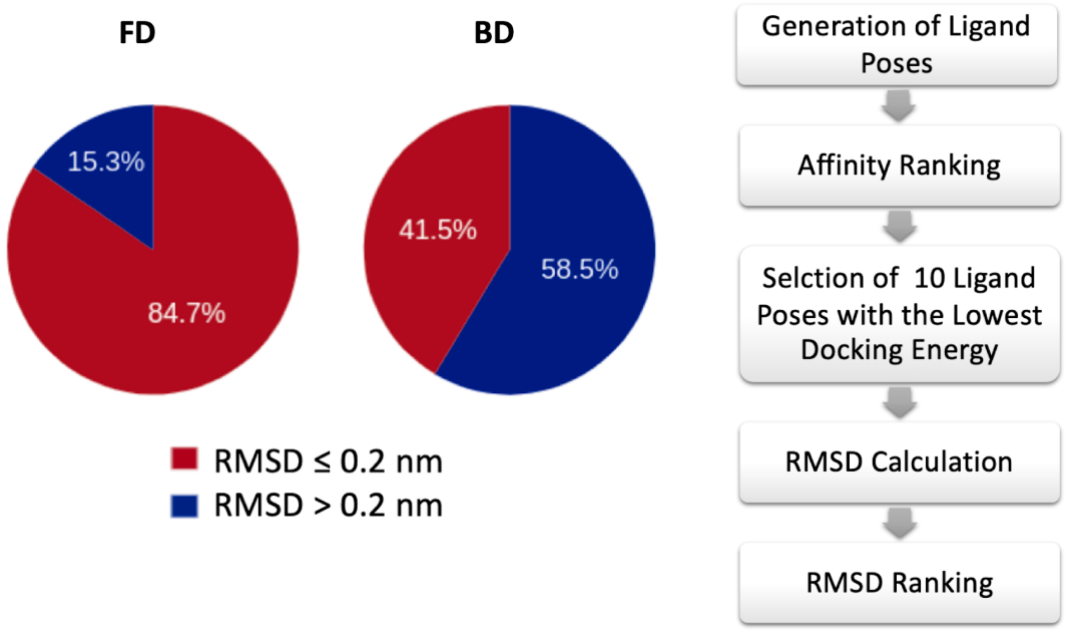


Figure 23. Percentage of poses generated with RMSD lesser and greater than 0.2 compared the original pose using the box and blind method. The comparison has been made considering the poses with the best RMSD value between first 10 with the maximum affinity value. Results of FD methods are shown in the pie chart on the left and results of BD method are shown in the pie chart on the right. Red indicates the correct configuration found and blue the wrong ones. An accurate ligand configuration has been discovered in 84.7% of cases using FD method and 41.5% using BD method.

The distribution of RMSD values is shown in Figure 24. Compared to the previous case, the number of conformations belonging to the first interval has increased. The maximum frequency is 149 in boxes and 73 in BD method.

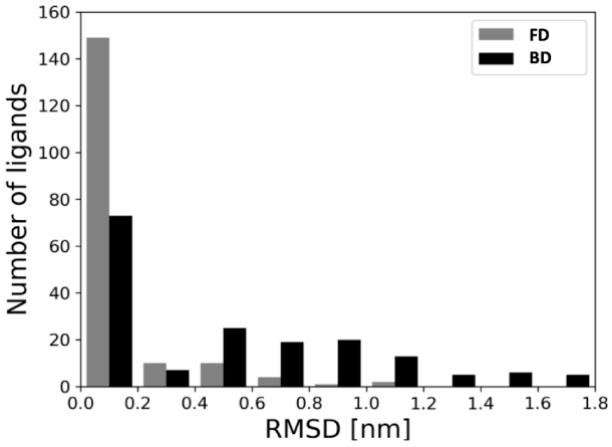


Figure 24. Histogram shows the distribution of the computed RMSD value considering the ten poses generated with greater affinity. In both method most of the generated configurations belong at the range between 0 and 0.2 nm. They are 149 in FD method and 73 in BD method.

In Figure 25 the relationship between the improving in performance with the increase in the number of poses considered is exhibited. The affinity ranking has been made including 1, 3, 5, 10, 20, 40, 70, 100 ligands configurations. The curve significantly rises from 1 to 5 poses considered. This suggests that Vina does not always attribute the highest affinity to the pose that closest to the real pose of the ligand.

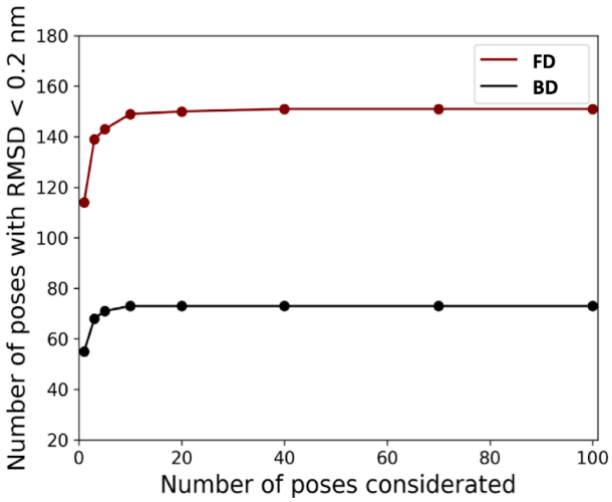


Figure 25. The curve underlines how the performances vary as the number of poses to consider calculating the RMSD increases. The greatest growth of the curve occurs when it goes from 0 to 5 in FD and BD method. Red curve refers the FD method, black curve to BD method.

4.3.2 MM/GBSA Rescoring

The evaluation of Vina's pose made considering the RMSD as principal parameter for the ranking, pointed out that the pose with the lowest RMSD does not always correspond to the one

with greater affinity. In order to address this issue, a rescoring protocol on outgoing poses has been accomplished. Binding free energy has been calculated using the MM/GBSA method on the first ten protein-ligand complex whose affinity is the highest calculated by Vina. The number of protein-ligand complex has been set to ten for the reason that this threshold shows the greatest improvement of finding the correct pose, as highlighted in Figure 18 and Figure 25. In Figure 26 the rescoring results for Hsp90α and HIV-1 PR structures are represented. In 75.7% of cases the greatest binding affinity estimated with MM/GBSA corresponds to ligand pose with a RMSD lesser than 0.2 nm compared the ligand in the crystal structure. The percentage of correct poses prediction is higher comparing the results with the case in which the ranking is done with the affinity estimated by Vina.In HIV-1 PR the performance is about the same as that obtained by considering the docking affinities.

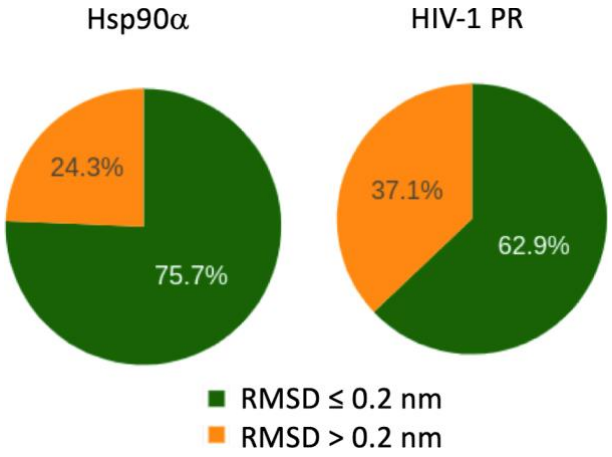


Figure 26 .Rescoring results. The percentage of the poses generated with RMSD lesser and greater than 0.2 nm. compared the original pose using FD method are shown for Hsp90α and HIV-1 PR structures. Results of Hsp90α are shown in the pie chart on the left and results of HIV-1 PR are shown in the pie chart on the right. Green indicates the correct configuration found and orange the wrong ones. An accurate ligand configuration has been discovered in 75.7% of cases in Hsp90α and 62.9% in HIV-1 PR.

4.4 Discussion

Protein–ligand docking is a powerful tool in drug discovery to predict binding modes and affinities of ligand. The blind docking is a common strategy employed when the binding site of a target is unknown[119,120]. However, blind docking requires great computational resources and the results obtained are often not accurate[104,121]. Several methods have been developed to overcome these critical issues. The strategies proposed in the literature usually employ specific software to find the active site in the target and then docking the ligands into the discovered binding site[104,122,123,124]. These alternative methods show that the blind docking results without

the aid of methodologies that identify the binding site are characterized by a very low percentage of successful runs[121]. However, if the real binding site is not included in the boxes in which the docking is carried out, the procedure could lead to an incorrect result. A different approach examined in protein-peptide docking is to perform a blind docking and then a re-docking focused in the binding site, proving that the binding site information reduces searching space drastically obtaining a fast e more precise results in re-docking. The performance of all the docking methods improved during the re-docking study[125]. Despite a slight increase in accuracy, the computation cost of this procedure is higher and the issue of finding the right binding site may remain, resulting in a useless the re-docking procedure if the binding pocket is not matched. In this work, a new docking procedure has been performed on Hsp90α and HIV-1 PR protein-ligand complexes employing a protocol that performs multiple smaller docking run along the surface of the protein in order to compare the performance with the blind method. The outputs of multiple box docking algorithm are several poses of ligand which bind protein in different sites. The main goal of performing the docking procedure dividing the protein surface in small boxes, is to allow the scan of the entirely protein surface which intrinsically implies the correct finding of the real binding site. The parameter used for validation is the RMSD, since the experimental ligand conformation coordinates are known. The RMSD has been calculated between the ligand conformations predicted by the Fragmented Docking (FD) and Blind Docking (BD) algorithm and the experimental one. The values obtained in FD and BD methods have been compared. A first comparison has been carried out between the RMSD values obtained considering only the ligand pose with the lowest docking energy. As can be seen from the pie charts shown above, the performances obtained with the FD method are better than the BD method in both protein-ligand complexes analyzed. The method of preforming the research in several boxes that translate on the protein has proven to be a winning strategy in finding the binding site on the protein and the relative ligand configuration.

A second analysis has been accomplished in order to evaluate the performance of Vina's scoring function, as done before in literature[126,127]. Docking algorithms, like Vina, have been developed with the goal to screening large numbers of compounds in a reasonable time, for this reason, they use approximate scoring functions. It is reasonable to hypothesis that the best pose, namely the one with the minimum RMSD from experimental data, is not always the first in the rank. Indeed, it has become general opinion that docking results should be improved by means of more rigorous post-docking processing strategies[128,129,127]. In literature several works have

evaluated the ability of molecular mechanics combined with the Poisson–Boltzmann surface area (MM-PBSA) and molecular mechanics combined with generalized Born surface area to predict binding affinities and compared the accuracy of these predictions to that of docking scores. A study on Protein kinase proved that the correlation between calculated binding free energies with MM/GBSA and MM/PBSA methods and experimental values is higher than using docking scores[130]. However, the results depend on the dataset used. For example, in a set of β-Amyloid Cleaving Enzyme 1 (BACE-1) the re-scoring docking poses using MM-GBSA did not improve the correlation with experimental affinities due to the ligand dataset[131]. BACE-1 ligands dataset consisted of macrocycles which have multiple flexible bonds that generate a large conformational space and for this reason require a more accurate MM/GBSA protocol[131]. In recent literature, the performance of MM/PBSA and MM/GBSA rescoring have been evaluated in protein-protein docking showing that MM/GBSA may be a good choice for predicting the binding affinities and identifying correct binding structures[132]. Following the aforementioned rescoring suggestion, in this work the rescoring strategy adopted involves the use of the MM/GBSA method. The binding affinity has been recalculated for the top ten poses with the greater docking energy. The results deriving from this rescoring show that MM/GBSA is able most of the time to attribute least energy to the pose with an RMSD lower than 0.2 nm. In literature, the same procedure of MM/GBSA rescoring has been applied to the three top scored docking poses showing that the results have improved compared the case in which only the best scored docking pose is considered[133,134]. The method is a good compromise between efficiency and speed since it has been applied on minimized protein-ligand complexes. Further challenge could be accomplished by calculating the affinity in MD simulated complexes or with more rigorous methods such as FEP with the disadvantage of a more computational time.

5 Conclusion and Future Developments

The docking protocol developed in this thesis has demonstrated its ability to address the protein-ligand docking where the binding sites are not known a priori. The algorithm is a simple and fast method that shifts the searching box on the protein surface in order to predict the correct binding sites. The presented results have demonstrated how this methodology improves the accuracy both in terms of binding site identification and of RMSD of the lowest energy docked pose with respect to the experimental solution. Moreover, the idea of performing a rescoring on the results generated by the docking algorithm employing methods of estimating the free binding energy proved to be a good solution for improvement.

In general, molecular docking has proven to be an efficient method to predict the experimental ligand conformation adopted in the target binding site. However, standard docking protocols employ only one structure to represent the protein, neglecting the changes in the geometry of the binding pocket induced by the ligand binding. Standard docking is carried out between the protein and the ligand extracted from the same crystallography. If a ligand is docked on another structure of the same protein (cross-docking) the results may be not always optimal. In these cases, it may happen that the internal cavity of the binding pocket does not have enough space to accommodate another ligand. Possible solutions provide to keep protein flexible during the docking procedure, but this requires higher computational effort. Future studies are needed to overcome this limit, by explicitly considering the protein conformational plasticity, which is sometimes a key point to estimate the drug selectivity/specificity for a protein target.

Acknowledgment

First of all, I would like to thank my supervisor Prof. Marco Agostino Deriu for introducing me to the Molecular Modelling research field and giving me the opportunity to be part of it. I am very grateful for giving me all the helps I needed and for giving me the special opportunity to work in the Computational Biophysics Group of the Dalle Molle Institute for Artificial Intelligence.

I would like to show my gratitude to my Co-Supervisors Dr. Gianvito Grasso for his guidance, support and suggestions.

Thanks to Prof. Andrea Danani of Computational Biophysics Group of the Dalle Molle Institute for Artificial Intelligence for accepting me in his research group.

A special thanks goes to Stefano and Filip for their valuable advice and for having supported me at every phase of the work.

Finally, I must express my gratitude to my family for having always believed in me and for providing me with unfailing support throughout my years of study and lifetime.

Arianna

References

1. Lindsay, M. A. Target discovery. *Nat. Rev. Drug Discov.* **2**, 831–838 (2003).
2. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
3. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997). *Adv. Drug Deliv. Rev.* **46**, 3–26 (2001).
4. Hughes, J., Rees, S., Kalindjian, S. & Philpott, K. Principles of early drug discovery. *Br. J. Pharmacol.* **162**, 1239–1249 (2011).
5. Bacilieri, M. & Moro, S. Ligand-Based Drug Design Methodologies in Drug Discovery Process: An Overview. *Curr. Drug Discov. Technol.* **3**, 155–165 (2006).
6. Surabhi, S. & Singh, B. COMPUTER AIDED DRUG DESIGN: AN OVERVIEW. *J. Drug Deliv. Ther.* **8**, 504–509 (2018).
7. Veselovsky, A. & Ivanov, A. Strategy of Computer-Aided Drug Design. *Curr. Drug Target - Infectious Disord.* **3**, 33–40 (2003).
8. Śledź, P. & Caflisch, A. Protein structure-based drug design: from docking to molecular dynamics. *Curr. Opin. Struct. Biol.* **48**, 93–102 (2018).
9. Krieger, E., Nabuurs, S. B. & Vriend, G. Homology Modeling. in 509–523 (2005). doi:10.1002/0471721204.ch25.
10. Batool, M., Ahmad, B. & Choi, S. A Structure-Based Drug Discovery Paradigm. *Int. J. Mol. Sci.* **20**, 2783 (2019).
11. Aparoy, P., Kumar Reddy, K. & Reddanna, P. Structure and Ligand Based Drug Design Strategies in the Development of Novel 5- LOX Inhibitors. *Curr. Med. Chem.* **19**, 3763–3778 (2012).
12. Macalino, S. J. Y., Gosu, V., Hong, S. & Choi, S. Role of computer-aided drug design in modern drug discovery. *Arch. Pharm. Res.* **38**, 1686–1701 (2015).
13. Bernetti, M., Cavalli, A. & Mollica, L. Protein–ligand (un)binding kinetics as a new paradigm for drug discovery at the crossroad between experiments and modelling. *Medchemcomm* **8**, 534–550 (2017).
14. Copeland, R. A., Pompliano, D. L. & Meek, T. D. Drug–target residence time and its implications for lead optimization. *Nat. Rev. Drug Discov.* **5**, 730–739 (2006).

15. Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. II. *Berichte der Dtsch. Chem. Gesellschaft* **27**, 3479–3483 (1894).
16. Du, X. *et al.* Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods. *Int. J. Mol. Sci.* **17**, 144 (2016).
17. Chang, C.-E. & Gilson, M. K. Free Energy, Entropy, and Induced Fit in Host−Guest Recognition: Calculations with the Second-Generation Mining Minima Algorithm. *J. Am. Chem. Soc.* **126**, 13156–13164 (2004).
18. Gilson, M. K. & Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 21–42 (2007).
19. de Azevedo Jr., W. & Dias, R. Computational Methods for Calculation of Ligand-Binding Affinity. *Curr. Drug Targets* **9**, 1031–1039 (2008).
20. Kairys, V., Baranauskiene, L., Kazlauskiene, M., Matulis, D. & Kazlauskas, E. Binding affinity in drug design: experimental and computational techniques. *Expert Opin. Drug Discov.* **14**, 755–768 (2019).
21. Bai, N., Roder, H., Dickson, A. & Karanicolas, J. Isothermal Analysis of ThermoFluor Data can readily provide Quantitative Binding Affinities. *Sci. Rep.* **9**, 2650 (2019).
22. Vuignier, K., Schappler, J., Veuthey, J.-L., Carrupt, P.-A. & Martel, S. Drug–protein binding: a critical review of analytical tools. *Anal. Bioanal. Chem.* **398**, 53–66 (2010).
23. Some, D. Light-scattering-based analysis of biomolecular interactions. *Biophys. Rev.* **5**, 147–158 (2013).
24. Patching, S. G. Surface plasmon resonance spectroscopy for characterisation of membrane protein–ligand interactions and its potential for drug discovery. *Biochim. Biophys. Acta - Biomembr.* **1838**, 43–55 (2014).
25. Nitsche, C. & Otting, G. NMR studies of ligand binding. *Curr. Opin. Struct. Biol.* **48**, 16–22 (2018).
26. Åqvist, J., Luzhkov, V. B. & Brandsdal, B. O. Ligand Binding Affinities from MD Simulations. *Acc. Chem. Res.* **35**, 358–365 (2002).
27. Gutiérrez-de-Terán, H. & Åqvist, J. Linear Interaction Energy: Method and Applications in Drug Design. in 305–323 (2012). doi:10.1007/978-1-61779-465-0_20.
28. Massova, I. & Kollman, P. A. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect. Drug Discov. Des.* **18**, 113–135 (2000).

29. Zhang, X., Perez-Sanchez, H. & C. Lightstone, F. A Comprehensive Docking and MM/GBSA Rescoring Study of Ligand Recognition upon Binding Antithrombin. *Curr. Top. Med. Chem.* **17**, 1631–1639 (2017).
30. Shivakumar, D. *et al.* Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J. Chem. Theory Comput.* **6**, 1509–1519 (2010).
31. Khavrutskii, I. V. & Wallqvist, A. Improved Binding Free Energy Predictions from Single-Reference Thermodynamic Integration Augmented with Hamiltonian Replica Exchange. *J. Chem. Theory Comput.* **7**, 3001–3011 (2011).
32. Kästner, J. Umbrella sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 932–942 (2011).
33. Do, P.-C., Lee, E. H. & Le, L. Steered Molecular Dynamics Simulation in Rational Drug Design. *J. Chem. Inf. Model.* **58**, 1473–1482 (2018).
34. Limongelli, V. Ligand binding free energy and kinetics calculation in 2020. *WIREs Comput. Mol. Sci.* (2020) doi:10.1002/wcms.1455.
35. Wang, W., Wang, J. & Kollman, P. A. What determines the van der Waals coefficient ? in the LIE (linear interaction energy) method to estimate binding free energies using molecular dynamics simulations? *Proteins Struct. Funct. Genet.* **34**, 395–402 (1999).
36. Kumari, R., Kumar, R. & Lynn, A. g_mmpbsa —A GROMACS Tool for High-Throughput MM-PBSA Calculations. *J. Chem. Inf. Model.* **54**, 1951–1962 (2014).
37. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **10**, 449–461 (2015).
38. Wang, E. *et al.* End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chem. Rev.* **119**, 9478–9508 (2019).
39. Genheden, S., Nilsson, I. & Ryde, U. Binding Affinities of Factor Xa Inhibitors Estimated by Thermodynamic Integration and MM/GBSA. *J. Chem. Inf. Model.* **51**, 947–958 (2011).
40. You, W., Tang, Z. & Chang, C. A. Potential Mean Force from Umbrella Sampling Simulations: What Can We Learn and What Is Missed? *J. Chem. Theory Comput.* **15**, 2433–2443 (2019).
41. Chelli, R., Marsili, S., Barducci, A. & Procacci, P. Generalization of the Jarzynski and Crooks nonequilibrium work theorems in molecular dynamics simulations. *Phys. Rev. E* **75**, 050101 (2007).
42. Wong, C. F. Steered molecular dynamics simulations for uncovering the molecular mechanisms of drug dissociation and for drug screening: A test on the focal adhesion kinase.

J. Comput. Chem. **39**, 1307–1318 (2018).

43. Hernndez-Santoyo, A., Yair, A., Altuzar, V., Vivanco-Cid, H. & Mendoza-Barrer, C. Protein-Protein and Protein-Ligand Docking. in *Protein Engineering - Technology and Application* (InTech, 2013). doi:10.5772/56376.
44. Gupta, M., Sharma, R. & Kumar, A. Docking techniques in pharmacology: How much promising? *Comput. Biol. Chem.* **76**, 210–217 (2018).
45. Ferreira, L., dos Santos, R., Oliva, G. & Andricopulo, A. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **20**, 13384–13421 (2015).
46. Sterling, T. & Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
47. Bolton, E. E., Wang, Y., Thiessen, P. A. & Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. in 217–241 (2008). doi:10.1016/S1574-1400(08)00012-1.
48. Wishart, D. S. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–D672 (2006).
49. Hanwell, M. D. *et al.* Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminform.* **4**, 17 (2012).
50. Evans, D. A. History of the Harvard ChemDraw Project. *Angew. Chemie Int. Ed.* **53**, 11140–11145 (2014).
51. Prieto-Martínez, F. D., Arciniega, M. & Medina-Franco, J. L. Acoplamiento Molecular: Avances Recientes y Retos. *TIP Rev. Espec. en Ciencias Químico-Biológicas* **21**, (2018).
52. Pagadala, N. S., Syed, K. & Tuszynski, J. Software for molecular docking: a review. *Biophys. Rev.* **9**, 91–102 (2017).
53. Alogheli, H., Olanders, G., Schaal, W., Brandt, P. & Karlén, A. Docking of Macrocycles: Comparing Rigid and Flexible Docking in Glide. *J. Chem. Inf. Model.* **57**, 190–202 (2017).
54. Ewing, T. J., Makino, S., Skillman, A. G. & Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided. Mol. Des.* **15**, 411–28 (2001).
55. Miller, M. D., Kearsley, S. K., Underwood, D. J. & Sheridan, R. P. FLOG: A system to select ?quasi-flexible? ligands complementary to a receptor of known three-dimensional structure. *J. Comput. Aided. Mol. Des.* **8**, 153–174 (1994).
56. Novič, M., Tibaut, T., Anderluh, M., Borišek, J. & Tomašič, T. The Comparison of Docking Search Algorithms and Scoring Functions. in 99–127 doi:10.4018/978-1-5225-0115-2.ch004.

57. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by Simulated Annealing. *Science (80-.).* **220**, 671–680 (1983).

58. Yang, X.-S. Genetic Algorithms. in *Nature-Inspired Optimization Algorithms* 77–87 (Elsevier, 2014). doi:10.1016/B978-0-12-416743-8.00005-1.

59. Liu, M. & Wang, S. MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *J. Comput. Aided. Mol. Des.* **13**, 435–51 (1999).

60. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–91 (2009).

61. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267**, 727–48 (1997).

62. Dias, R. & de Azevedo Jr., W. Molecular Docking Algorithms. *Curr. Drug Targets* **9**, 1040–1047 (2008).

63. Leone, V., Marinelli, F., Carloni, P. & Parrinello, M. Targeting biomolecular flexibility with metadynamics. *Curr. Opin. Struct. Biol.* **20**, 148–154 (2010).

64. Li, J., Fu, A. & Zhang, L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdiscip. Sci. Comput. Life Sci.* **11**, 320–328 (2019).

65. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3**, 935–949 (2004).

66. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004).

67. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* NA-NA (2009) doi:10.1002/jcc.21334.

68. Abagyan, R., Totrov, M. & Kuznetsov, D. ICM?A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **15**, 488–506 (1994).

69. Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. A Fast Flexible Docking Method using an Incremental Construction Algorithm. *J. Mol. Biol.* **261**, 470–489 (1996).

70. Schnecke, V. & Kuhn, L. A. Database screening for HIV protease ligands: the influence of binding-site conformation and representation on ligand selectivity. *Proceedings. Int. Conf. Intell. Syst. Mol. Biol.* 242–51 (1999).

71. Korb, O., Stützle, T. & Exner, T. E. Empirical Scoring Functions for Advanced Protein−Ligand

Docking with PLANTS. *J. Chem. Inf. Model.* **49**, 84–96 (2009).

72. Grosdidier, A., Zoete, V. & Michielin, O. EADock: Docking of small molecules into protein active sites with a multiobjective evolutionary optimization. *Proteins Struct. Funct. Bioinforma.* **67**, 1010–1025 (2007).
73. Corbeil, C. R., Williams, C. I. & Labute, P. Variability in docking success rates due to dataset preparation. *J. Comput. Aided. Mol. Des.* **26**, 775–786 (2012).
74. McGann, M. FRED and HYBRID docking performance on standardized datasets. *J. Comput. Aided. Mol. Des.* **26**, 897–906 (2012).
75. Pang, Y.-P., Perola, E., Xu, K. & Prendergast, F. G. EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. *J. Comput. Chem.* **22**, 1750–1771 (2001).
76. Venkatachalam, C. M., Jiang, X., Oldfield, T. & Waldman, M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph. Model.* **21**, 289–307 (2003).
77. Jain, A. N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *J. Med. Chem.* **46**, 499–511 (2003).
78. Baxter, C. A., Murray, C. W., Clark, D. E., Westhead, D. R. & Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **33**, 367–82 (1998).
79. Welch, W., Ruppert, J. & Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **3**, 449–462 (1996).
80. Wu, G., Robertson, D. H., Brooks, C. L. & Vieth, M. Detailed analysis of grid-based molecular docking: A case study of CDOCKER?A CHARMm-based MD docking algorithm. *J. Comput. Chem.* **24**, 1549–1562 (2003).
81. Tietze, S. & Apostolakis, J. GlamDock: Development and Validation of a New Docking Tool on Several Thousand Protein−Ligand Complexes. *J. Chem. Inf. Model.* **47**, 1657–1672 (2007).
82. Mizutani, M. Y., Tomioka, N. & Itai, A. Rational Automatic Search Method for Stable Docking Models of Protein and Ligand. *J. Mol. Biol.* **243**, 310–326 (1994).
83. Thomsen, R. & Christensen, M. H. MolDock: A New Technique for High-Accuracy Molecular Docking. *J. Med. Chem.* **49**, 3315–3321 (2006).
84. Zsoldos, Z., Reid, D., Simon, A., Sadjad, S. B. & Johnson, A. P. eHiTS: A new fast, exhaustive flexible ligand docking system. *J. Mol. Graph. Model.* **26**, 198–212 (2007).
85. Muegge, I., Martin, Y. C., Hajduk, P. J. & Fesik, S. W. Evaluation of PMF scoring in docking

weak ligands to the FK506 binding protein. *J. Med. Chem.* **42**, 2498–503 (1999).

86. Bhm, H.-J. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput. Aided. Mol. Des.* **6**, 593–606 (1992).

87. Xie, Z.-R. & Hwang, M.-J. An interaction-motif-based scoring function for protein-ligand docking. *BMC Bioinformatics* **11**, 298 (2010).

88. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V & Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided. Mol. Des.* **11**, 425–45 (1997).

89. Fan, H. *et al.* Statistical potential for modeling and ranking of protein-ligand interactions. *J. Chem. Inf. Model.* **51**, 3078–92 (2011).

90. Krammer, A., Kirchhoff, P. D., Jiang, X., Venkatachalam, C. M. & Waldman, M. LigScore: a novel scoring function for predicting binding affinities. *J. Mol. Graph. Model.* **23**, 395–407 (2005).

91. Gohlke, H., Hendlich, M. & Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **295**, 337–56 (2000).

92. Yin, S., Biedermannova, L., Vondrasek, J. & Dokholyan, N. V. MedusaScore: An Accurate Force Field-Based Scoring Function for Virtual Drug Screening. *J. Chem. Inf. Model.* **48**, 1656–1662 (2008).

93. Reulecke, I., Lange, G., Albrecht, J., Klein, R. & Rarey, M. Towards an integrated description of hydrogen bonding and dehydration: decreasing false positives in virtual screening with the HYDE scoring function. *ChemMedChem* **3**, 885–97 (2008).

94. DeWitte, R. S. & Shakhnovich, E. I. SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *J. Am. Chem. Soc.* **118**, 11733–11744 (1996).

95. Gehlhaar, D. K. *et al.* Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **2**, 317–24 (1995).

96. Das, S., Krein, M. P. & Breneman, C. M. Binding Affinity Prediction with Property-Encoded Shape Distribution Signatures. *J. Chem. Inf. Model.* **50**, 298–308 (2010).

97. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **47**, 2977–2980 (2004).

98. Preto, J. & Gentile, F. Assessing and improving the performance of consensus docking strategies using the DockBox package. *J. Comput. Aided. Mol. Des.* **33**, 817–829 (2019).

99. Preto, J. *et al.* Molecular Dynamics and Related Computational Methods with Applications to Drug Discovery. in 267–285 (2018). doi:10.1007/978-3-319-76599-0_14.

100. Combes, J. M., Duclos, P. & Seiler, R. The Born-Oppenheimer Approximation. in *Rigorous Atomic and Molecular Physics* 185–213 (Springer US, 1981). doi:10.1007/978-1-4613-3350-0_5.

101. Nelder, J. A. & Mead, R. A Simplex Method for Function Minimization. *Comput. J.* **7**, 308–313 (1965).

102. Verbeke, J. & Cools, R. The Newton-Raphson method. *Int. J. Math. Educ. Sci. Technol.* **26**, 177–193 (1995).

103. Hassan, N. M., Alhossary, A. A., Mu, Y. & Kwoh, C.-K. Protein-Ligand Blind Docking Using QuickVina-W With Inter-Process Spatio-Temporal Integration. *Sci. Rep.* **7**, 15451 (2017).

104. Ghersi, D. & Sanchez, R. Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. *Proteins Struct. Funct. Bioinforma.* **74**, 417–424 (2009).

105. Prodromou, C., Roe, S. M., Piper, P. W. & Pearl, L. H. A molecular clamp in the crystal structure of the N-terminal domain of the yeast Hsp90 chaperone. *Nat. Struct. Biol.* **4**, 477–482 (1997).

106. Prodromou, C. & Pearl, L. Structure and Functional Relationships of Hsp90. *Curr. Cancer Drug Targets* **3**, 301–323 (2003).

107. Garrido, C., Gurbuxani, S., Ravagnan, L. & Kroemer, G. Heat Shock Proteins: Endogenous Modulators of Apoptotic Cell Death. *Biochem. Biophys. Res. Commun.* **286**, 433–442 (2001).

108. Deeks, S. G. HIV-1 Protease Inhibitors. *JAMA* **277**, 145 (1997).

109. Brik, A. & Wong, C.-H. HIV-1 protease: mechanism and drug discovery. *Org. Biomol. Chem.* **1**, 5–14 (2003).

110. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma.* **54**, (2016).

111. Sanner, M. F. Python: a programming language for software integration and development. *J. Mol. Graph. Model.* **17**, 57–61 (1999).

112. Nissink, J. W. M. *et al.* A new test set for validating predictions of protein-ligand interaction. *Proteins Struct. Funct. Bioinforma.* **49**, 457–471 (2002).

113. Cole, J. C., Murray, C. W., Nissink, J. W. M., Taylor, R. D. & Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins Struct. Funct. Bioinforma.* **60**, 325–332 (2005).

114. Gabb, H. A., Jackson, R. M. & Sternberg, M. J. E. Modelling protein docking using shape complementarity, electrostatics and biochemical information 1 1Edited by J. Thornton. *J. Mol. Biol.* **272**, 106–120 (1997).

115. Zaheer-ul-Haq, Halim, S. A., Uddin, R. & Madura, J. D. Benchmarking docking and scoring protocol for the identification of potential acetylcholinesterase inhibitors. *J. Mol. Graph. Model.* **28**, 870–882 (2010).

116. Lape, M., Elam, C. & Paula, S. Comparison of current docking tools for the simulation of inhibitor binding by the transmembrane domain of the sarco/endoplasmic reticulum calcium ATPase. *Biophys. Chem.* **150**, 88–97 (2010).

117. Miller, B. R. *et al.* MMPBSA.py : An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **8**, 3314–3321 (2012).

118. Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium. *J. Phys. Chem.* **100**, 19824–19839 (1996).

119. Hetényi, C. & van der Spoel, D. Blind docking of drug-sized compounds to proteins with up to a thousand residues. *FEBS Lett.* **580**, 1447–1450 (2006).

120. Hetényi, C. & van der Spoel, D. Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci.* **11**, 1729–1737 (2009).

121. Brown, W. M. & Vander Jagt, D. L. Creating Artificial Binding Pocket Boundaries To Improve the Efficiency of Flexible Ligand Docking. *J. Chem. Inf. Comput. Sci.* **44**, 1412–1422 (2004).

122. Blaszczyk, M. *et al.* Modeling of protein–peptide interactions using the CABS-dock web server for binding site search and flexible docking. *Methods* **93**, 72–83 (2016).

123. Lin, Y.-F. *et al.* MIB: Metal Ion-Binding Site Prediction and Docking Server. *J. Chem. Inf. Model.* **56**, 2287–2291 (2016).

124. Liu, Y. *et al.* CB-Dock: a web server for cavity detection-guided protein–ligand blind docking. *Acta Pharmacol. Sin.* **41**, 138–144 (2020).

125. Agrawal, P. *et al.* Benchmarking of different molecular docking methods for protein-peptide docking. *BMC Bioinformatics* **19**, 426 (2019).

126. Gaillard, T. Evaluation of AutoDock and AutoDock Vina on the CASF-2013 Benchmark. *J. Chem. Inf. Model.* **58**, 1697–1706 (2018).

127. Quiroga, R. & Villarreal, M. A. Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. *PLoS One* **11**, e0155183 (2016).

128. Rastelli, G. & Pinzi, L. Refinement and Rescoring of Virtual Screening Results. *Front. Chem.* **7**, (2019).

129. Palacio-Rodríguez, K., Lans, I., Cavasotto, C. N. & Cossio, P. Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. *Sci. Rep.* **9**, 5142 (2019).

130. Sun, H., Li, Y., Tian, S., Xu, L. & Hou, T. Assessing the performance of MM/PBSA and MM/GBSA methods. 4. Accuracies of MM/PBSA and MM/GBSA methodologies evaluated by various simulation protocols using PDBbind data set. *Phys. Chem. Chem. Phys.* **16**, 16719–16729 (2014).

131. El Khoury, L. *et al.* Comparison of affinity ranking using AutoDock-GPU and MM-GBSA scores for BACE-1 inhibitors in the D3R Grand Challenge 4. *J. Comput. Aided. Mol. Des.* **33**, 1011–1020 (2019).

132. Chen, F. *et al.* Assessing the performance of the MM/PBSA and MM/GBSA methods. 6. Capability to predict protein–protein binding free energies and re-rank binding poses generated by protein–protein docking. *Phys. Chem. Chem. Phys.* **18**, 22129–22139 (2016).

133. Sun, H. *et al.* Assessing the performance of MM/PBSA and MM/GBSA methods. 5. Improved docking performance using high solute dielectric constant MM/GBSA and MM/PBSA rescoring. *Phys. Chem. Chem. Phys.* **16**, 22035–22045 (2014).

134. Rastelli, G., Rio, A. Del, Degliesposti, G. & Sgobba, M. Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *J. Comput. Chem.* NA-NA (2009) doi:10.1002/jcc.21372.

Supporting Information

Table S 1. RMSD value between the coordinates of experimentally ligand conformation and the predicted by the algorithm for Hsp90α. First column refers the name of PDB protein-ligand complex and the ligand selected; the second and third column refer RMSD value considering the predicted ligand pose with best affinity value in FD and BD, respectively; the fourth and fifth column refer RMSD value considering the poses with the best RMSD value between first ten with the maximum affinity in FD and BD, respectively;

PDB - LIGAND	FD_RMSD_1[nm]	BD_RMSD_1[nm]	FD_RMSD_10[nm]	BD_RMSD_10[nm]
1A4H-GDM	0,037	1,548	0,036	1,095
1AMW-ADP	0,342	0,342	0,179	0,216
1BGQ-RDC	0,051	0,331	0,049	0,331
1UY6-PU3	0,069	0,066	0,067	0,066
1UY7-PU4	0,438	0,439	0,152	0,146
1UY8-PU5	0,405	0,405	0,128	0,141
1UY9-PU6	0,336	0,338	0,124	0,127
1UYC-PU7	0,343	0,345	0,067	0,126
1UYD-PU8	0,322	0,388	0,164	0,311
1UYE-PU9	0,421	0,462	0,169	0,356
1UYF-PU1	0,695	0,684	0,194	0,338
1UYH-PU0	0,662	0,666	0,066	0,083
1UYI-PUZ	0,402	0,685	0,143	0,505
1UYM-PU3	0,086	0,084	0,084	0,084
1YC1-4BC	0,097	0,090	0,097	0,090
1YC4-43P	0,037	0,037	0,036	0,036
1YET-GDM	0,028	2,543	0,028	1,289
2BRC-CT5	0,158	0,498	0,157	0,191
2BYH-2D7	0,115	0,373	0,110	0,373
2BYI-2DD	0,094	0,389	0,093	0,388
2CCS-4BH	0,060	0,501	0,059	0,501
2CCU-2D9	0,139	0,126	0,139	0,125
2FWY-H64	0,167	0,855	0,142	0,542
2FWZ-H71	0,097	0,781	0,091	0,621
2IWS-NP4	0,028	0,278	0,028	0,278

2IWU-NP5	0,049	0,277	0,048	0,277
2IWX-M1S	0,053	0,259	0,052	0,258
2QG0-A94	0,726	1,783	7,368	0,546
2QG2-A91	0,517	0,514	0,198	0,212
2UWD-2GG	0,105	0,104	0,512	0,514
2VCI-2GJ	0,169	2,241	0,056	0,103
2VCJ-2EQ	0,113	2,091	0,168	1,424
2VWC-BC2	0,061	1,529	0,106	1,505
2WEQ-GDM	0,071	1,221	0,061	1,395
2WI1-ZZ2	0,407	0,377	0,070	1,221
2WI3-ZZ3	0,303	0,304	0,074	0,078
2WI4-ZZ4	0,524	0,546	0,109	0,268
2WI5-ZZ5	0,532	0,533	0,184	0,331
2WI6-ZZ6	0,058	0,373	0,356	0,532
2WI7-2KL	0,795	0,696	0,057	0,372
2XDK-XDK	0,414	0,312	0,280	0,513
2XDL-2DL	0,033	0,553	0,398	0,274
2XDX-WOE	0,395	0,489	0,033	0,480
2XHR-C0P	0,426	0,495	0,238	0,280
2XHT-C0Y	0,035	0,103	0,110	0,426
2XHX-T5M	0,088	0,091	0,034	0,103
2XJG-XJG	0,024	0,026	0,088	0,090
2XJX-XJX	0,303	0,303	0,024	0,025
2XX4-13I	0,045	0,617	0,131	0,153
2XX5-13N	0,647	0,595	0,044	0,414
2YGA-GDM	0,047	0,762	0,093	0,414
2YGE-GDM	0,063	0,068	0,046	0,762
2YGF-GDM	0,739	1,651	0,062	0,067
2YI0-YI0	0,058	0,816	0,633	1,138
2YI5-YI5	0,513	0,397	0,058	0,777
2YI7-BZ8	0,116	0,576	0,107	0,329
2YJW-YJW	0,510	0,383	0,114	0,519

2YJX-YJX	0,029	0,028	0,243	0,383
2YK9-YK9	0,745	0,124	0,028	0,027
2YKB-YKB	0,086	0,088	0,123	0,124
2YKC-YKC	0,030	0,025	0,086	0,088
2YKE-YKE	0,032	0,034	0,025	0,025
2YKI-YKI	0,022	1,111	0,031	0,033
3B24-B2J	0,205	0,204	0,020	0,806
3B25-B2K	0,058	0,051	0,197	0,203
3B27-B2T	0,043	0,043	0,053	0,050
3BM9-BXZ	0,048	0,600	0,032	0,043
3BMY-CXZ	0,182	0,390	0,048	0,574
3D0B-SNX	0,316	0,316	0,182	0,389
3FT5-MO8	0,066	0,234	0,062	0,070
3NMQ-7PP	0,182	0,396	0,066	0,080
3OW6-MEX	0,042	0,045	0,078	0,366
3OWB-BSM	0,055	0,053	0,041	0,045
3OWD-MEY	0,745	0,068	0,054	0,052
3QDD-94M	0,068	0,332	0,067	0,068
3QTF-05S	0,054	0,054	0,068	0,332
3R4M-WOE	0,414	0,438	0,051	0,054
3R91-06H	0,038	0,037	0,065	0,219
3R92-06J	0,041	0,502	0,032	0,036
3RKZ-06T	0,050	0,055	0,040	0,502
3VHA-VHA	0,020	0,018	0,050	0,054
3VHC-VHC	0,031	0,926	0,020	0,017
4AS9-4QS	0,061	0,073	0,031	0,569
4B7P-9UN	0,366	0,367	0,060	0,072
4BQG-50Q	0,147	0,437	0,148	0,149
4CE1-7FK	0,077	0,255	0,144	0,436
4CE2-BO5	0,053	0,273	0,076	0,254
4CE3-L4V	0,104	0,234	0,053	0,272
4CWF-H05	0,411	0,411	0,035	0,233

4CWN-6LV	0,706	0,759	0,184	0,191
4CWO-T62	0,125	0,126	0,038	0,057
4CWP-TV2	0,697	0,695	0,124	0,126
4CWQ-W2D	0,036	0,032	0,027	0,029
4CWR-HAJ	0,723	0,724	0,031	0,032
4CWS-G3R	0,151	0,078	0,015	0,348
4CWT-IK9	0,251	0,253	0,151	0,078
4EEH-HH6	0,676	0,629	0,105	0,164
4EFT-EFT	0,237	0,233	0,032	0,033
4EFU-EFU	0,115	0,115	0,047	0,232
4EGH-0OY	0,066	0,083	0,114	0,114
4EGK-RDC	0,026	0,329	0,064	0,083
4FCQ-2N6	0,490	0,494	0,025	0,329
4FCR-0TM	0,034	0,354	0,048	0,493
4LWE-FJ2	0,854	0,365	0,034	0,354
4LWH-FJ5	0,049	0,035	0,113	0,365
4LWI-FJ6	0,092	0,098	0,038	0,035
4O04-2Q8	0,131	1,838	0,091	0,097
4O07-FGH	0,033	1,742	0,130	0,765
4O09-2R6	0,036	1,878	0,031	0,772
4O0B-2QA	0,032	0,810	0,034	0,792
4W7T-3JC	0,652	0,304	0,031	0,708
4XIP-40W	0,087	0,712	0,121	0,124
4XIT-40Z	0,099	0,581	0,086	0,640
5FNC-IEE	0,368	0,378	0,087	0,580
5FND-IQ5	0,378	0,322	0,298	0,229
5FNF-TQL	0,712	0,715	0,373	0,321

Table S 2 RMSD value between the coordinates of experimentally ligand conformation and the predicted by the algorithm for HIV-1 PR. First column refers the name of PDB protein-ligand complex and the ligand selected; the second and third column refer RMSD value considering the predicted ligand pose with best affinity value in FD and BD, respectively; the fourth and fifth column refer RMSD value considering the poses with the best RMSD value between first ten with the maximum affinity in FD and BD, respectively;

PDB - LIGAND	**FD_RMSD_1[nm]**	**BD_RMSD_1[nm]**	**FD_RMSD_10[nm]**	**BD_RMSD_10[nm]**
1A94-0Q4	1,029	1,5038	1,029	1,079
1A9M-U0E	1,216	1,1161	0,158	1,116
1AAQ-PSI	0,144	1,0876	0,117	0,570
1AJV-NMB	0,091	0,6471	0,091	0,639
1AJX-AH1	0,156	2,1078	0,119	1,827
1B6K-PI5	0,060	0,0525	0,060	0,052
1B6L-PI4	0,055	0,9737	0,054	0,881
1BDL-IM1	1,119	1,1058	1,106	1,105
1BDQ-IM1	0,962	0,9560	0,078	0,120
1BDR-IM1	0,104	0,1390	0,104	0,139
1BV7-XV6	0,868	0,7191	0,712	0,478
1BV9-XV6	0,461	1,9819	0,461	1,199
1C6Y-MK1	0,119	0,1198	0,119	0,119
1C70-L75	0,101	0,0713	0,091	0,071
1D4H-BEH	0,040	1,0941	0,040	1,070
1D4I-BEG	1,077	0,0297	0,030	0,029
1DIF-A85	0,093	0,1103	0,093	0,110
1DMP-DMQ	0,062	1,6519	0,062	1,554
1EBY-BEB	0,034	1,0183	0,034	0,948
1EBZ-BEC	0,078	0,0787	0,078	0,078
1EC0-BED	1,053	0,4611	0,023	0,411
1EC1-BEE	0,054	0,9135	0,054	0,833
1G2K-NM1	0,131	1,0711	0,131	1,071
1G35-AHF	0,056	0,7779	0,056	0,777
1HBV-GAN	1,107	0,3733	0,123	0,373
1HEG-PSI	2,088	0,6046	0,608	0,574
1HIV-1ZK	1,092	0,5346	0,363	0,534

1HPO-UNI	1,095	1,0690	0,063	0,852
1HPX-KNI	0,220	0,2190	0,220	0,081
1HSG-MK1	0,051	0,0478	0,051	0,047
1HTE-G23	0,700	0,7001	0,128	0,135
1HTF-G26	0,655	0,6559	0,655	0,641
1HVI-A77	0,086	0,0906	0,086	0,090
1HVJ-A78	0,091	0,0910	0,091	0,091
1HVK-A79	0,039	0,0399	0,039	0,039
1HVL-A76	0,056	0,1023	0,056	0,102
1HVR-XK2	0,104	1,2842	0,104	1,265
1HVS-A77	0,101	0,0748	0,101	0,074
1HXW-RIT	0,360	1,0514	0,106	0,632
1IIQ-0ZR	0,575	1,1699	0,468	0,980
1IZI-Q50	1,192	1,1900	0,146	0,137
1KZK-JE2	0,078	1,0639	0,078	0,885
1MES-DMP	0,918	0,9187	0,038	0,043
1MET-DMP	0,083	0,0771	0,081	0,077
1MEU-DMP	0,114	0,0774	0,112	0,077
1MRW-K57	0,076	0,0294	0,042	0,029
1MRX-K57	0,028	0,0284	0,028	0,028
1MSM-JE2	0,079	0,6209	0,079	0,620
1MSN-JE2	0,028	2,0794	0,028	1,271
1MTR-PI6	0,576	0,7233	0,328	0,588
1ODY-LP1	1,246	2,1549	0,668	0,818
1OHR-1UN	0,077	0,5936	0,026	0,593
1PRO-A88	0,037	1,9590	0,037	0,916
1QBR-XV6	0,567	0,5706	0,551	0,570
1QBS-DMP	0,920	1,1523	0,057	0,964
1SBG-IM1	0,387	0,3803	0,125	0,124
1SDT-MK1	0,076	0,0845	0,076	0,084
1SDU-MK1	0,078	0,0782	0,078	0,078
1SDV-MK1	0,080	0,0803	0,039	0,080

1SGU-MK1	0,370	0,1136	0,266	0,113
1SH9-RIT	1,269	0,8536	0,428	0,474
1TCX-IM1	0,050	0,0961	0,050	0,096
1VIJ-BAY	0,497	0,4649	0,447	0,254
1VIK-BAY	1,487	0,1883	0,189	0,188
1W5V-BE3	0,029	1,0181	0,029	1,018
2AOD-2NC	0,154	1,1826	0,154	0,709
2BBB-HH1	0,105	2,0355	0,105	1,664
2BPV-1IN	0,382	0,3828	0,152	0,216
2BPX-MK1	0,066	0,0678	0,066	0,067
2BPY-3IN	0,172	0,1726	0,162	0,172
2BQV-A1A	0,100	1,0110	0,067	0,863
2CEJ-1AH	0,073	0,9293	0,073	0,912
2HB3-GRL	0,086	0,0839	0,086	0,083
2I0A-MUI	0,143	1,0915	0,143	0,479
2I0D-MUT	0,601	0,4569	0,092	0,347
2I4D-QFI	0,154	0,1344	0,139	0,134
2I4U-DJR	0,069	0,9847	0,069	0,858
2I4V-DJR	0,061	1,0066	0,061	0,486
2I4X-KGQ	0,771	0,7988	0,193	0,779
2O4L-TPV	1,169	0,5825	0,128	0,576
2O4N-TPV	0,087	0,6831	0,087	0,670
2O4P-TPV	0,125	0,7930	0,125	0,736
2P3B-3TL	1,319	1,8149	0,376	1,814
2PQZ-G0G	0,164	0,6561	0,085	0,448
2PSU-MUU	0,127	1,9719	0,100	1,653
2PSV-MUV	0,521	1,8822	0,161	1,183
2PWC-G3G	0,118	0,1584	0,116	0,158
2PWR-G4G	0,918	0,8773	0,174	0,741
2Q54-MU1	1,088	0,4916	0,149	0,161
2Q55-MU0	0,048	0,0869	0,048	0,086
2Q5K-AB1	0,076	1,0620	0,076	0,798

2Q64-1UN	0,080	0,0803	0,080	0,080
2QHY-MZ1	0,047	0,3782	0,047	0,378
2QHZ-MZ2	0,112	1,5432	0,112	1,543
2QI0-MZ3	0,147	0,1646	0,117	0,164
2QI1-MZ4	0,080	0,0831	0,080	0,083
2QI3-MZ5	0,341	0,3386	0,336	0,169
2QI4-MZ6	0,148	2,5044	0,119	1,571
2QI5-MZ7	0,093	1,7563	0,093	1,707
2QI6-MZ8	0,122	0,1055	0,062	0,105
2QNP-QN2	0,066	1,0016	0,066	0,548
2QNQ-QN3	0,697	0,7183	0,495	0,696
2R38-G4G	0,096	0,9014	0,096	0,882
2R3T-G4G	0,092	0,0926	0,092	0,092
2R3W-G3G	0,111	0,1115	0,110	0,111
2R43-G3G	0,111	0,1556	0,111	0,155
2UPJ-U02	0,145	0,1433	0,145	0,143
2UY0-HV1	0,098	0,4785	0,098	0,478
2ZGA-YDP	1,826	0,8701	0,815	0,764
3AID-ARQ	0,778	0,4357	0,411	0,159
3BGB-LJG	0,881	0,8310	0,105	0,657
3BGC-LJH	0,927	0,6303	0,089	0,255
3BXS-DRS	0,536	0,4522	0,062	0,452
3CKT-YDP	0,103	0,1026	0,103	0,102
3EKP-478	0,448	0,5138	0,135	0,505
3EKQ-ROC	1,083	1,3264	0,064	1,184