Università degli Studi di Cagliari

PHD DEGREE
Cycle: XXXII

An enhanced-sampling MD-based protocol for molecular docking

Scientific Disciplinary Sectors:
FIS07, FIS03

PhD Student: Andrea Basciu

Coordinator of the PhD Programme: Prof. Paolo Ruggerone

Supervisors: Prof. Paolo Ruggerone and Dr. Attilio V. Vargiu

Final exam. Academic year 2018/2019
Thesis defence: January-February 2020 Session

Abstract

Understanding the binding of small molecules to proteins in atomistic detail is key for drug design. Molecular docking is a widely used computational method to mimic ligand-protein association *in silico*. However, predicting the conformational changes occurring in proteins upon ligand binding is still a major challenge. Ensemble docking approaches address this issue by considering a set of different conformations of the protein obtained either experimentally or from computer simulations, e.g. from molecular dynamics. However, bound-like (holo) structures prone to host (the correct) ligands are generally poorly sampled by standard molecular dynamics simulations of the unbound (apo) protein. In order to address this limitation, we introduce a computational approach based on metadynamics simulations called *ensemble docking with enhanced sampling of pocket shape* (*EDES*) that allows holo-like conformations of proteins to be generated by exploiting only their apo structures. This is achieved by defining a set of collective variables able to sample different shapes of the binding site, ultimately mimicking the steric effect due to the ligand. In this work, we assessed the method on re-docking and cross-docking calculations. In the first case, we selected three different protein targets undergoing different extents of conformational changes upon binding and, for each of them, we docked the experimental ligand conformation into an ensemble of receptor structures generated by EDES. In the second case, in the context of a blind docking challenge, we generated the 3D structures of a set of different ligands of the same receptor and docked them into a set of EDES-generated conformations of that receptor. In all cases, for both re-docking and cross-docking experiments, our protocol generates a significant fraction of structures featuring a low RMSD from the experimental holo geometry of the receptor. Moreover, ensemble docking calculations using those conformations yielded in almost all cases to native-like poses among the top-ranked ones. Finally, we also tested an improved EDES recipe on a further target, known to be extremely challenging due to its extended binding region and to the large extent of conformational changes accompanying the binding of its ligands.

Contents

List of Figures

List of Tables

List of Abbreviations

AMD	Accelerated Molecular Dynamics
BE-META	Bias-Exchange Metadynamics
BS	Binding Site
CA	Cluster Analysis
CAAD	Computer-Aided Drug Design
CV	Collective Variable
DOF	Degree of Freedom
EA	Evolutionary Algorithms
FES	Free Energy Surface
FEL	Free Energy Landscape
FF	Force Fields
IFD	Induced Fit Docking
MC	Monte Carlo
MD	Molecular Dynamics
META	Metadynamics
MR	Molecular Recognition
NMR	Nuclear Magnetic Resonance
PBC	Periodic Boundary Conditions
QM	Quantum Mechanical
RCS	Relaxed Complex Scheme
RMSD	Root Mean Square Deviation
ROG	Radius of Gyration
SA	Simulated Annealing
SBDD	Structure-Based Drug Design
SF	Scoring Function
SO	Swarm Optimisations
VDW	Van der Waals
VS	Virtual Screening
WT-META	Well-Tempered Metadynamics

1.1 Background and motivation

Highly specific and tightly regulated intermolecular interactions among biological molecules, collectively named "*molecular recognition (MR) events*", are the key tools by which cells control virtually all of their processes [1, 2]. Indeed, the generation of biological material such as nucleic acids, proteins and lipids, as well as intracellular communication and metabolism, all rely on the interaction between two or more biological molecules. A special class of MR events is the one involving the interaction of small molecules (*ligands*) with specific regions (*binding sites*, BSs) of target macromolecules, such as proteins (the vast majority of interactors [3]) and nucleic acids. In particular, in the case of proteins, BSs are usually small clefts on protein's surface featuring a certain degree of physico-chemical affinity for the ligands.

Quantitative understanding of ligand-receptor MR events is not only of great importance for basic research in life sciences but also a key prerequisite for modern drug design efforts. Indeed, despite the mechanisms by which ligands exert their actions once bound to the receptors are manifold, a common requisite is that the association is characterised by a good binding affinity. Experimental [3, 4] and computational [2, 5] studies revealed that the interactions underlying MR events between small ligands and their receptors are most often non-covalent (also named weak or non-bonding interactions). Due to these weak interactions, ligand-receptor association is usually characterised by a transient nature with specific timescales for the initiation and duration of the binding [6], linked to the intrinsic dynamical behaviour of macromolecules. Indeed, it has been recognised since long [7] that proteins are flexible molecular machines that are best described considering them not in a single static structure, but in an ensemble of energetically accessible (and interchanging) conformations representing all possible functional states [8–13]. Conformational transitions among those states, usually initiated and/or stabilised by ligand interactions and occurring also prior to intimate binding, can be considered at the basis of MR events, as they lead to the physico-chemical correspondences between partner's exposed surfaces necessary for their complexation [2, 14].

Concerning the extent of binding-induced conformational changes, it has been

shown that is extremely case dependent [15] (Figure 1.1), being sensitive to the nature of both partners. Most often, rearrangements are limited to the BS region and involve only sidechain reorientations and/or small hinge movements of the receptor around the ligand [16]. However, several examples can be found of MR events accompanied by medium-scale distortions involving loops and/or confined secondary structure variations and even large-scale motions among (sub)domains (e.g., hinge-bending or shear motions) leading to an extended reorganisation of the whole protein structure [16–19]. Cooperative proteins bearing multiple putative binding sites on different subunits [20, 21] are a typical example of the latter case, where ligand binding on a subunit alters the affinity of other monomers for the ligand via allosteric (extended) conformational changes. Examples of such systems, where the binding is accompanied by extended conformational rearrangements, are considered for example in refs. [22–25]. Moreover, another common feature in ligand-protein binding is the partial collapse of the BS upon binding, leading to the compaction of the putative site around the ligand [26, 27]. Although not completely general but still case dependent and linked to receptor's nature and ligand's size, this behaviour has been observed in several classes of pharmaceutically relevant proteins (such as kinases [28] transferases [29] and synthases [30]).

The ability to address dynamic conformational changes taking place in proteins

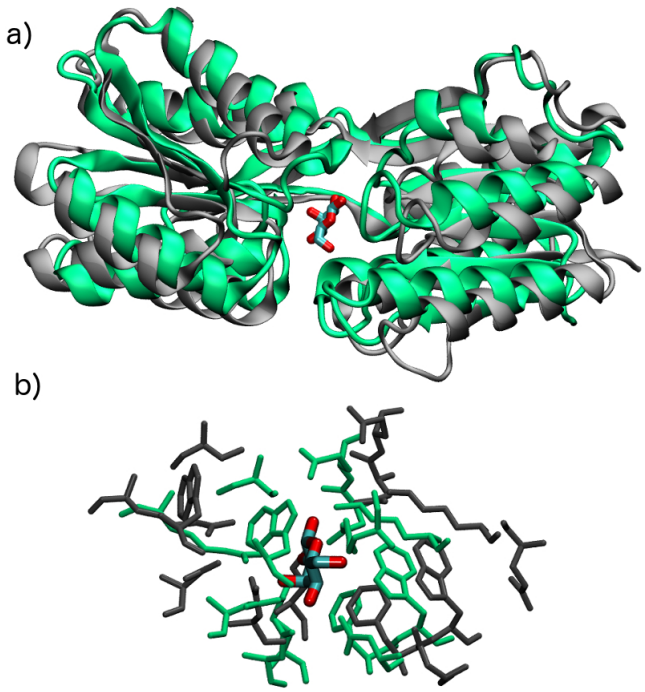

Figure 1.1: Conformational changes accompanying the binding of small ligands. In the picture, an example of conformational changes accompanying the binding of a small ligand is represented. Unbound (apo) and bound (holo) protein conformations are shown as ribbons colored black and green, respectively; the ligand is shown as stick colored by element. (a) Overall conformational rearrangements of the protein. (b) Detailed view of the local rearrangements occurring at the binding site.

upon ligand binding has been one of the most relevant and challenging issues in drug design since early times. Indeed, it is not known in advance which conformation flexible targets will adopt in response to the binding of a particular ligand, or how to design such a ligand for an unknown receptor conformation. To cope with this issue, several studies attempted to classify protein motions and their structural diversity (including ligand binding among the triggering events), leading sometimes to the creation of public databases, such as the "protein structural change upon ligand binding database" (PSCDB) [31] and the "database of protein conformational diversity in the native state" (CoDNaS 2.0) [32].

Experimental methods such as X-ray crystallography [33], cryo-electron microscopy (cryo-EM) [34] and nuclear magnetic resonance (NMR) [35] have been essential in elucidating the extent and the types of conformational changes accompanying MR events involving proteins, nucleic acids and other biological macromolecules. In the last decades, a huge number of protein structures have been determined by these methods in both their unbound (apo) and bound (holo) conformation, most of which available in the Protein Data Bank (PDB) [36], allowing for extensive and systematic studies on the distribution and types of conformational changes occurring along with MR events [37–39]. Among the above-mentioned experimental techniques, X-ray crystallography has been the most used technique for structure determination. While providing very high spatial resolutions for samples of virtually any molecular weight, crystallography furnishes in most cases only a static picture of the systems, sometimes also containing structural artefacts due to the intrinsic nature of this approach, requiring to build the 3D structure based on an electron density map. For this reason, in case the artefacts interest the putative binding region or this region presents a marked flexibility crucial for the binding event, X-ray structures alone may not be suitable to unveil the details of ligand association processes. Furthermore, due to the requirement of very pure and highly-concentrated samples, their production and crystallisation can be very challenging, time expensive and costly, particularly for large-sized proteins or proteins requiring a particular environment to stay in a near-native conformation (such as membrane proteins). On the other hand, NMR studies still require a pure and highly-concentrated sample, but they have the advantage of accessing to macromolecules in solution, thus catching the dynamic nature of the ensemble of different conformations produced. Finally, cryo-EM requires a much smaller amount of sample which does not need to be crystallised, making this technique less sensitive to impurities than X-ray crystallography. Moreover, the rapid freezing of the sample, prior to the cryo-EM analysis, increases the chances of the macromolecule to stay near its native-state with respect to crystallisation. However, its spatial resolution is in general lower than that normally achieved with X-ray crystallography and NMR (although resolution boosted up in the last decade), and the need for a high signal to noise ratio makes cryo-EM applicable only to large proteins [40–44].

As mentioned, while providing a solid framework for understanding the structural determinants of MR, these methods still present limitations that can prevent them to be used in specific situation [7, 45–49]. Examples of these cases are studies requesting to unveil details of the motions accompanying ligand-target interactions (such as the record of very short-live conformational metastable states) or to study receptors for

which the experimental treatment is difficult (e.g. because of difficulties to crystallise them in conditions resembling the physiological one) [46–48, 50, 51]. Moreover, the exploration of different conformations in experimental structures is generally limited and biased toward (often just a few) known ligand-receptor complexes, which impacts on the chemical diversity of putative lead compounds that can be studied experimentally [7, 45–47, 49].

As a complement to experiments, a plethora of theoretical/computational methods aiming to predict in silico the structures of ligand-receptor complexes have bloomed up in the recent years [52]. Some of these computational approaches, such as molecular docking [53, 54], molecular dynamics (MD) [55, 56] and Monte Carlo (MC) simulations [57, 58], have become key tools of modern structural biology [2, 54, 56, 59, 60] allowing to overcome some of the intrinsic limitations underlying experimental settings. In particular, these methods are used (often combined together) to attempt to describe the dynamics of ligand-receptor MR events, going beyond the static or limitedly flexible view furnished by atomic resolution experimental methods [22, 49, 57, 59, 61–63]. For example, advanced MD simulations are able to access at a relatively low computational cost non-equilibrium conformers of proteins, ligands, and their complexes (e.g. intermediate states of complexation) [2, 22, 24, 59, 64–67]. This being relevant as also poorly visited (high energy) (meta)stable states of (macro)molecules, e.g. proteins, may be crucial for ligand association [2].

Among the above-mentioned methods, molecular docking is by far the most used computational tool in the field of computer-aided drug design (CADD)[2, 15, 54, 67–73] with the aim to retrieve ligand-receptor complex structure starting from the unbound conformations of the binding partners. In the following, we will restrict our analysis to protein-ligand docking, in which a ligand (typically a small and low molecular weight organic compound), binds to a protein receptor. In this case, docking aims at mimicking the binding process between the two partners (i) to predict if they form a stable complex and (ii) to reproduce the conformation of the bound complex. In particular, scope of docking algorithms is to characterise the so-called binding pose (hereafter "pose") of a ligand into a receptor, identifying ligand's location and conformation within receptor's putative binding pocket [72, 74, 75].

Typically, protein-ligand docking calculations consist of two main steps, which although formally distinguishable, are intimately related to each other: (i) the searching (or sampling) step, consisting in the generation of the binding poses and (ii) the scoring step, consisting in the evaluation of the probability of that ligand/receptor complex to really occur and be stable, which is therefore ranked accordingly. Clearly, successful docking calculations strongly depend on how the conformational rearrangements of the binding partners are accounted as even small structural changes can seriously affect the quality of results also due to the limitations of the scoring step [2, 26, 72, 73]. Indeed, several strategies to treat partner's flexibility also prior to docking calculations exist, such as ensemble docking [54, 74, 76–79], in which a set of different receptor and/or ligand geometries, experimentally or computationally obtained, are used in the docking run, thus accounting for partners' structural plasticity prior to the docking step.

Within the ensemble-docking framework, here we propose a new approach called *Ensemble Docking with Enhanced sampling of pocket Shape* (EDES) [22] exploiting

relatively short metadynamics simulations [80] of the apo protein to generate a set of structures resembling its holo conformations [79], without exploiting a priori information about the holo structure. The method has been tested on three different target proteins paradigms of systems undergoing different extent of conformational changes. Moreover, its performance when coupled to a strategy to take care of ligand flexibility has also been addressed. Finally, we tested an improved EDES recipe also on a further protein, the adelynate kinase enzyme, presenting a very extended binding region composed of two (sub)pockets and undergoing the largest structural rearrangements upon binding among all the other targets addressed in this thesis [25, 81–83].

1.2 Thermodynamics of ligand binding

In a simple reversible interaction, the formation of the complex (PL) between a protein (P) and a ligand (L) can be described as a two-step process [2, 84–86] (Eq. 1.1):

$$P + L \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} PL \tag{1.1}$$

where k_{on} ($M^{-1}s^{-1}$) and k_{off} (s^{-1}) are the kinetic rate constants accounting, respectively, for the association and the dissociation of the binding partners. At equilibrium, where the rate of association equals that of dissociation ($k_{on}[P][L] = k_{off}[PL]$, with [X] representing the concentration of the "X" chemical species), the kinetic properties of a system are studied introducing a new set of quantities, collectively named "equilibrium constants". Ligand association, in particular, is typically studied by means of the binding and dissociation constants, respectively K_b and K_d, defined as in eq. 1.2:

$$K_b = \frac{[PL]}{[P][L]} = \frac{k_{on}}{k_{off}} = \frac{1}{K_d} \tag{1.2}$$

where $[PL]$, $[P]$ and $[L]$ are, the concentrations of respectively the complex (PL), the protein (P) and the ligand (L). From a complementary point of view, binding events can also be explained by the laws of thermodynamics. According to thermodynamics, a spontaneous reaction, such as ligand binding, at constant temperature and pressure (as it usually happens in a biological context) and at equilibrium, will occur only if accompanied by a negative change in the Gibbs free energy of the system (often named "free energy of binding"). Moreover, if the complexation occurs under standard conditions (1 atm pressure, 298 K temperature, and $[L] = [P]$ =1 M), the change in Gibbs free energy takes the name of "*standard* free energy of binding" (ΔG°), which can be easily related to the dissociation constant K_d using the relationship in eq. 1.3:

$$\Delta G^\circ = -RT \ln(K_d) \tag{1.3}$$

where R is the gas universal constant and T the absolute temperature. Because of the link between K_d and ΔG°, the stability of any ligand-receptor complex is said to be determined by the (negative) magnitude of ΔG°.

In typical cases, however, ligand binding occurs under non-standard conditions, so that ΔG° doesn't represent the true free energy of binding. Still a strong relationship exists between ΔG° and ΔG as shown in eq. 1.4, which indicates that the free energy of binding under non-standard conditions (ΔG) can be calculated on the basis of the one in standard conditions (ΔG°) corrected by an additional term containing the "reaction quotient" Q, defined as $[PL]/[P][L]$. Indeed, at equilibrium, the "reaction quotient" takes the value K_d reducing thus eq. 1.4 to eq. 1.3.

$$\Delta G = \Delta G^\circ + RT \ln(Q) \tag{1.4}$$

The free energy of binding can also be expressed in a different form, as in eq. 1.5, in which the molecular determinants of the interaction are highlighted:

$$\Delta G = \Delta H - T\Delta S \tag{1.5}$$

Equation 1.5 shows that the change in Gibbs free energy depends on two different contributions, an enthalpic (ΔH) and an entropic (-$T\Delta S$) term, reflecting the different kinds of interactions involved in binding events [1, 2, 84, 86–88]. The enthalpic contribution reflects the specificity and strengths of direct protein-ligand interactions, such as electrostatics, van der Waals and polarisation-induced forces. Moreover, it also accounts for solvent-mediated interactions between the binding partners and for the ones between the partners and the solvent [1, 2, 89]. Concerning water-mediated interactions, such as hydrogen-bonds (H-Bonds), it has been showed they can be of extreme relevance for ligand-protein complexation [88, 90–92]. In particular, as discussed for example by Chen et al. [93], H-Bonds facilitate ligand binding via at least two different mechanisms: (i) by establishing direct ligand-receptor interactions [94] and (ii) by displacing protein-bound (structural) water molecules into the bulk solvent [95]. On the other hand, the entropic term ΔS, in its simplest description, is considered as a measure of the dynamics of the overall system during the complexation. Changes in the binding entropy usually reflect the loss of motion associated to the formation of the complex and, in particular, the loss of available degrees of freedom of the binding partners and of the (structural) water molecules at the binding interface [89]. Indeed this term is usually modelled as made up of different contributions, as for example done by Perozzo and coworkers in ref. [88] and shown in eq. 1.6:

$$\Delta S = \Delta S_{solv} + \Delta S_{conf} + \Delta S_{r/t} \tag{1.6}$$

where ΔS_{solv} accounts for solvent release upon binding, ΔS_{conf} represents the conformational entropy change associated to structural re-organizations of partners before intimate binding, and $\Delta S_{r/t}$ refers to the change (loss) of translational and rotational degrees of freedom of the binding partners when the complex is formed from the partners free in solution. Although entropic contributions are usually refereed as unfavourable for ligand binding, in some cases they, and particularly conformational entropy contributions, have been discovered to favour the complexation [96]. As both enthalpic and entropic terms ultimately drive the binding, they are collectively called the "thermodynamic signature" of the association reaction [1]. Their knowledge is thus much more informative than just the value of ΔG value alone. Indeed, being able to distinguish between the two contributions is greatly helpful to shed light

on the molecular determinants of ligand-receptor association allowing for a direct comparison between computational and experimental methods for the two terms separately. Moreover, their knowledge can also improve lead optimisation steps in drug design (both experimentally and computationally) by giving hints on the types and extents of lead modifications needed to improve binding affinity between the compound under study and the target protein(s) [1]. Although different in origin, these contributions are highly correlated [1, 94], as, for example, partners' conformational flexibility (contributing to the entropic term) reflects the strength of non-covalent interactions governing the binding (and giving rise to an enthalpic contribution). Moreover, the strong connection between the two can also be retrieved in other two different phenomena, the enthalpy-entropy compensation effect, where a favourable change (increase) in system's enthalpy during complexation is often balanced by a corresponding unfavourable entropy change [94, 97] and hydrophobic interactions, often seen as the paradigm of water-mediated driving forces [90, 98–101].

Although several works in the last decades addressed this topic, helping to shed light on the intricate world of ligand binding energetics, a clear a complete picture is still missing. Among the reasons, the lack of strong theoretical models to quantitatively account for (de)solvation and entropic contributions being one of the most critical ones. However, as a rigorous treatment of the topic is out of the scope of this thesis, the interested reader is refereed to some recent literature on the field, such as refs. [2, 43, 89, 99].

1.3 Mechanisms of molecular recognition

Historically, the first mechanism to explain MR was the "lock-and-key model" proposed by Fisher in 1894 [102] and based on the matching between the binding surfaces of (virtually) rigid interacting partners (Figure 1.2). Thus, only the right ligand is able to enter and bind to receptor's BS, in the same way in which a key complements a lock. Although quite simplistic, a number of experimental evidences supported this model. For example, the binding determinants for the first X-ray resolved antibody-protein complex were understood by means of a static picture of the interacting partners [103]. Moreover, also the mechanism of action of chymotrypsin [104] was initially understood without considering protein flexibility. However, it was soon realised that for a deep understanding of biological interactions it was essential to account for the flexibility of (macro)molecules. Before introducing molecular plasticity in MR, it is interesting to mention that, as noted by Vogt et al. [10], using static models to understand ligand binding processes can still be appropriate in some cases, also in the framework of the overall intrinsic conformational flexibility of macromolecules. For instance, in the case of proteins exploring different conformations on a faster time scale compared to the time required for the binding process to occur, the equilibrium kinetics of the process would appear indistinguishable from the one associated to a static receptor, although eventual structural rearrangements can still occur after the binding event making, in that case, unsuitable the usage of a static model of the protein. This also explains why, as for the cases cited above, invoking protein plasticity is not always needed to understand shape fitting between ligand and receptor involving no or very minor conformational rearrangements, although the static picture of the binding event is rarely able to unveil the dynamics detail of

the interaction. The first mechanism of MR accounting in an explicit way protein plasticity was the "induced fit model", introduced more than 60 years after Fisher's seminal work by Koshland [105] to explain enzyme catalysis (Figure 1.2). According to this model, the increasing interactions between approaching partners induces complementary structural adaptations in both the receptor and the ligand, leading to optimal matching of their binding surfaces. The induced fit model has been extremely useful to elucidate the binding dynamics of a large number of protein classes [16, 106–108]. Furthermore, it gives a theoretical basis to understand the evidence of proteins binding multiple ligands [109]. However, it still does consider proteins as represented by a single binding-incompetent conformation in the absence of interacting ligands, and accounts for their flexibility only limitedly to their binding site(s) [2, 105, 110].

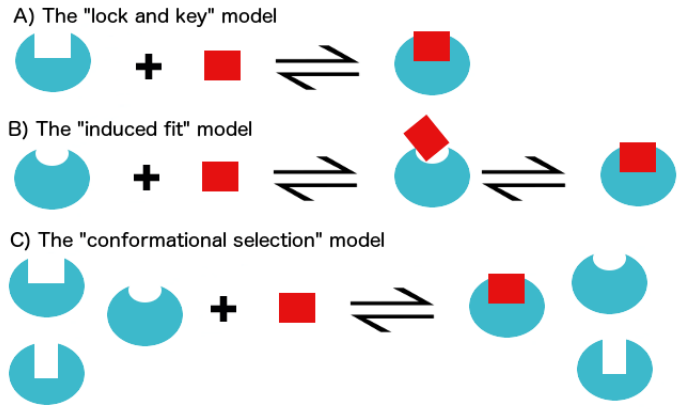

Figure 1.2: Main models of molecular recognition: lock and key, induced fit, and selective fit (a.k.a. conformational selection or dynamic fit). Proteins are represented in cyan, while red blocks indicate a generic partner (here a ligand), whose possible conformational transitions occurring upon binding are not reported for simplicity. In the lock and key model, no structural rearrangements occur upon ligand interaction and binding. In the induced fit model, the ligands bind the receptor in a weak conformation, inducing afterwards a conformational transition towards the tight conformation to maximise favourable interactions. In the selective fit model the receptor (and the ligand) adopts several substates in dynamical equilibrium, and only one of them is selected and stabilised upon ligand binding.

The "selective fit model" (aka "dynamic fit model" or "conformational selection model", Figure 1.2), at opposite, is based on the observation that proteins in their biological environment (also in the absence of a specific ligand) are intrinsically dynamic, assuming an ensemble of different (meta)stable states [2, 8, 110–112]. This model, firstly introduced in the 1960s by Straub and Szabolcsi [113] became widely accepted and broadly referenced by the scientific community around the 1980s, thanks, among the others, to the contributions by Nussinov and coworkers [114]. Notably, the model was then further improved and reshaped when Frauenfelder and coworkers [12] coupled the idea of protein intrinsic flexibility with the "free energy landscape (FEL) theory" [8, 12, 59, 115]. According to the FEL theory, the rich topology of proteins, coupled with the intrinsic nature of non-bonded interactions governing their dynamics, gives rise to a very complex free-energy landscape [13, 116–118] corresponding to

large ensembles of conformational (sub)states (hence conformers) coexisting with different population distributions in dynamic equilibrium [12, 119, 120]. Barriers comparable to the thermal energy will allow for the interconversion among different conformers, leading to a significant population also of (relatively) high-energy states. Indeed, several studies on protein-ligand interactions and enzymatic catalysis showed that, in certain cases, ligands can bind not only to the lowest energy state of a receptor (preferred by the protein in the absence of ligands) but also to higher energy conformations resembling its ligand-bound form, resulting in a subsequent population shift toward the latter conformational states [59, 115, 121]. The last decades have seen numerous efforts to combine the above models into a unified theory of protein-ligand binding. Models have been proposed in which a first step of weak ligand binding to the most complementary receptor geometry (according to the conformational selection theory) is followed by reciprocally induced conformational rearrangements so as to tighten the interactions [122, 123]. However, the relative importance of the two mechanisms involved in the unified model is still under debate, as it appears that the balance depends on the specific binding partners and on the details of the environment in which the binding takes place (e.g. on ligand concentration) [59, 123–125]. Finally, in the more recent years, also models trying to explain MR events only by means of the conformational selection theory, thus reconsidering the importance and predominance of an induced fit mechanism, have been proposed, as done in ref. [10].

1.4 Thesis outline

Aim of this thesis is to present the recently developed Ensemble Docking with Enhanced Sampling of pocket Shape (EDES) method aiming to sample holo-like structures of a protein receptor given its apo form. The method will be described in details, explaining the reasons why it can positively contribute to the scientific community working in the field of basic research in life sciences and in the field of rational drug design. Its performance will be tested in different scenarios, understanding its strength points but also pinpointing its weak spots. Conclusions will be then drawn highlighting possible directions for a further development of the method.

This thesis is organised as follows:

Chapter 2 is divided in two parts. In the first one, an overview of the theoretical background needed to understand EDES implementation will be given. The concepts of molecular dynamics simulations, metadynamics, docking calculations and cluster analysis will be addressed. In the second part, the details of EDES methodology will be discussed.

Chapter 3 deals with EDES performance in the context of re-docking and cross-docking calculations. In the first case, re-docking calculations on EDES-generated receptor conformations are performed for three different protein receptors being paradigms of targets undergoing different extents of conformational changes. In the second one, EDES performance in cross-docking calculations is addressed via our participation to the D3R Grand Challenge 4, a blind docking challenge where the participants are requested to predict the binding poses of a set of ligands on a protein

receptor for which only the amino acid sequence is known. Finally, we also tested an improved EDES recipe on a further target, known to be extremely challenging due to its extended binding region and to the large extent of conformational changes accompanying the binding of its ligands.

Chapter 4 draws some general conclusions on the results obtained, framing this work into a wider project. In this chapter we also discuss the perspectives of future works and possible directions of improvement of the method presented.

Computational Methods 2

Over the years, the approaches used to investigate molecular recognition events have changed. The development of new computer technology and of new computational techniques has considerably increased the accuracy of computational studies of biological systems. Computational methods have thus become an important tool in biomedicine. In particular, in addition to the experimental techniques, several computational methods have been applied at different stages of the drug-design processes [49, 126].

The work presented in this thesis is based on computational simulations. In this chapter an overview is given of the main computational methods used, such as classical molecular dynamics (MD), metadynamics, molecular docking and cluster analysis.

2.1 Molecular dynamics (MD) simulations

Molecular dynamics (MD) simulations have provided to be extremely useful to address problems with many degrees of freedom, and they have become one of the principal tool of the theoretical study of biological molecules and reactions. At a molecular level, life can be regarded as a complex network of chemical and physical interactions between (bio)chemical entities. In this context, MD simulations provide a practical and less expensive way to study the dynamic behaviour of (macro)molecules. Two main families of MD approaches have been developed over the years: classical MD and quantum mechanics (QM) MD.

In the QM approach, the quantum nature of bonds is explicitly considered. Electronic density functions are used to compute the dynamics of the valence electrons, while ions and inner electrons are still treated classically [55, 127–132]. However, quantum mechanics simulations need more computational resources and so they are usually used to simulate small portion of a biological system in the *ps* time scale, although approaches allowed to reach the *ns* scale [133, 134].

In classical MD simulations, on the other hand, atoms and bonds are treated classically: the electronic distributions of atoms are approximated considering fixed partial charges on them and by approximate models for polarisation effects, while

bonds are treated like springs, with empirical values for the force constants. In this way the dynamics of the system is defined by the laws of classical mechanics.

2.2 Classical MD

In classical MD, the potential energy of the system is computed by means of empirical energy functions which, together with the numerical parameters used, are collectively refereed as "force fields" (FF). Force fields are based upon the Born-Oppenheimer approximation and they are also refereed as atomistic models since the smallest particles involved in these functions are atoms, rather than electrons as in QM calculations [55, 56, 127]. The distances used in the calculations are only refereed to the positions of atoms (i.e. of nuclei) and electron motions are not taken into account explicitly but rather considered by an adjustment of the parameters used [135].

Within FF the interactions governing the structure and the evolution of the system are modelled by simple mathematical terms and, together with their functional form, FF contain a set of parameters for each type of atoms. More atom types than elements are present, since the properties of an atom is strongly influenced by its chemical environment. As an example, force fields include different parameters for a carbon atom according to its hybridisation (e.g. a *sp* carbon would have different parameters than a sp^2 one) [52, 135].

Several force fields exist and, although the functional form of many of them is very similar, the main differences arise in the parameters used and in the way in which the value of parameters are determined.

FF typically compute the conformational energy of a system as a summation of the contributes of the so called "bonded interactions" (covalent bonds) and of the "non-bonded interactions" (non covalent bonds), as described in equation 2.1:

$$E_{total} = E_{covalent} + E_{non-covalent} \tag{2.1}$$

where the two contributes are usually expressed as follows:

$$E_{covalent} = E_{bond} + E_{angle} + E_{dihedral} \tag{2.2}$$

$$E_{non-covalent} = E_{electrostatic} + E_{VDW} \tag{2.3}$$

The terms shown in equations 2.2 and 2.3 are used in MD simulations to calculate the force acting on every particle in the system as follows:

$$\mathbf{F}_i = -\nabla U_i, \quad i = 1, 2, ..., N \tag{2.4}$$

with U_i the potential energy associated to every of the N particles of the system. Considering all the contributes discussed so far, the functional form of the potential energy of a particle i due to the interactions with the other particles of the system reads:

$$U_i = \sum_k U_{i,k}^{bonds} + \sum_k U_{i,k}^{angles} + \sum_k U_{i,k}^{dihedrals} + \sum_j U_{i,j}^{electrostatic} + \sum_j U_{i,j}^{vdW} \tag{2.5}$$

Where, the index k indicates the particles covalently bound to the particle i while the index j refers to the non-bonded atoms. As clear from the above equation, the total potential energy is computed as a sum of pairwise interactions. It is worth keeping in mind that the functional form of equation 2.5 and the mathematical expression of the terms contained may slightly differ according to the specific force field considered. What is presented here the semi-empirical energy function used in the AMBER force field [136].

Bonded terms are associated with the covalently bound atoms (equation 2.6). Their functional forms is given by equations 2.7, 2.8 and 2.9:

$$U^{bonded} = U^{bonds} + U^{angles} + U^{dihedrals} \tag{2.6}$$

$$U^{bonds} = \sum_{bonds} K_b(r - r_{eq})^2 \tag{2.7}$$

$$U^{angles} = \sum_{angles} K_\theta(\theta - \theta_{eq})^2 \tag{2.8}$$

$$U^{dihedrals} = \sum_{dihedrals} K_\chi(1 + cos(n\phi - \phi_{eq})) \tag{2.9}$$

Before analysing each contribution, it is important to stress the difference between the experimentally obtained values and the FF-dependent parameters used in the equations above. The terms associated with the 3D structure of molecules, which are typically determined experimentally (by techniques such as X-ray crystallography and NMR spectroscopy) but also by means of high-level quantum calculations, are the bond lengths, r_{eq}, the valence angles, θ_{eq}, and the torsion or dihedral angles, ϕ_{eq}. The numerical values of the other parameters might vary for different force fields. As already pointed out, both experimentally determined terms and parameters depend not only on the chemical element involved but also on its chemical environment. For example, a C-C single bond may have r_{eq} =1.53 Å and K_b = 225 kcal/(mol Å^2) while a C=C group will have shorter bonds, e.g. r_{eq}=1.33 Å , and a stronger bond constant, e.g. K_b = 500 kcal/(mol Å^2). Note that the simple form of the equations above represents a compromise between accuracy and computational costs [52, 135].

The bond stretching term and the angular vibrations are treated harmonically, which is a good approximation for a system at room temperature, where both the bonds and the angles stay close to their equilibrium position. A more physically accurate treatment would require the usage of the Morse potential, resulting in an increase of the computational cost associated. However, the Morse potential would give significantly better results only at high temperatures (which are almost never used for biological systems) and for this reason it is rarely used.

The dihedral or torsion angles are associated with rotations around a covalent bond as shown in figure 2.1. This oscillatory terms contains the force constant K_χ, which indicates the height of the rotational barrier, the periodicity of the rotation n and the equilibrium angle ϕ_{eq}. It is important to note that rotating around a covalent bond can eventually result in a dramatic conformational energy change of the structures, due eventually induced steric clashes between different groups of the structures [135].

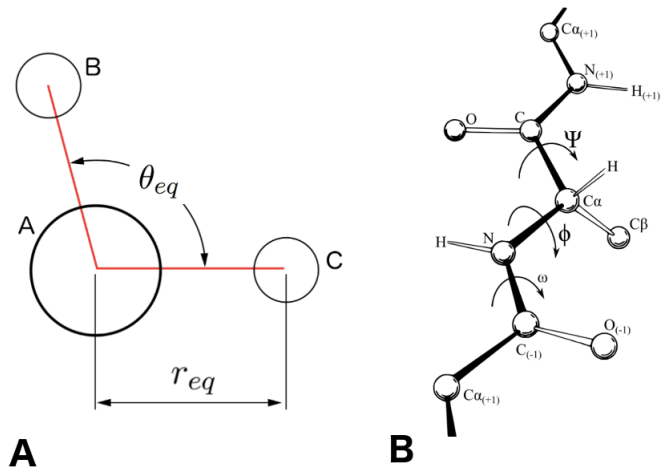

Figure 2.1: Bonded interactions: A) the equilibrium bond length r_{eq} and valence angle θ_{eq} for a simple molecule. B) a schematic representation of a protein with its three torsion angles ω, ϕ and ψ. Traditionally, torsion angles in proteins have different names according to the specific covalent bond to which they are refereed. They are all computed by eq. 2.9.

Non-bonded interactions are what drives almost every process of molecular recognition. Interestingly, although the proper physical treatment of these interaction can be very complex, it has been shown that accurate results can be obtained treating these interactions with a relatively simple mathematical model [130, 135].

Equations 2.10 and 2.11 show the simple models which are commonly used in almost all FF to compute these interactions:

$$U^{electrostatic} = \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_d r_{ij}} \tag{2.10}$$

$$U^{VDW} = \sum_{i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.11}$$

The electrostatic interaction (eq. 2.10) is modelled by means of a Coulombic term. This term involves an interaction between the partial charges q_i and q_j of atoms i and j divided by the distance between them and by the appropriate dielectric constant of the medium. Even if the system is globally neutral, the usage of partial charges is a way to include in the electrostatic term also the displacement of electronic densities, which leads to dipoles and multipoles.

Van der Waals interactions are described by the term shown in equation 2.11. This term is described by a semi-empirical potential refereed as the Lennard-Jones (figure 2.2). It is composed of two terms, one representing the repulsive interactions due to electron clouds overlapping, and the other one indicating the attractive interaction, typical of induced dipoles interactions. The repulsive term is proportional to r^{-12} and its effect is relevant only at very short distances. The attractive term is proportional to r^{-6} and it is negative, which indicates its favourable nature. The VDW term contains two parameters: ϵ_{ij} which indicates the magnitude of the interaction (i.e. the depth of the potential well) and σ_{ij}, indicating the (finite) distance at which $U^{electrostatic} = 0$.

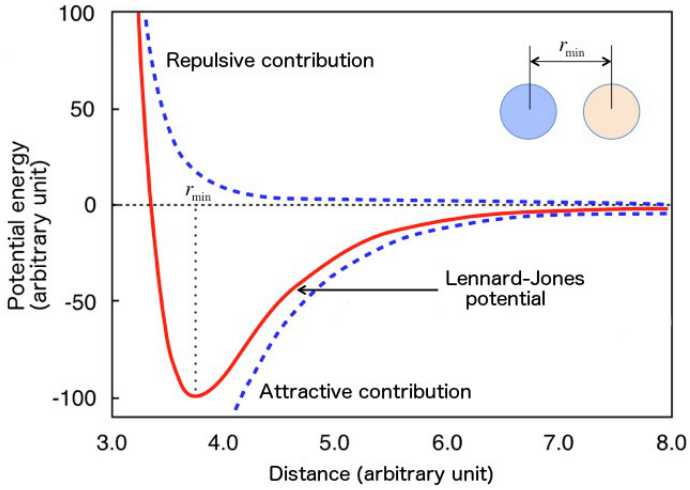

FIGURE 2.2: A representation of the Lennard-Jones 12-6 potential. Note the equilibrium distance r_{min} between two chemical entities in which the potential energy has a minimum.

Moreover, σ_{ij} is related to the distance between two interacting atoms at which interaction is most favourable: $r_{ij}^{min} = 2^{\frac{1}{6}}\sigma_{ij}$. Typically the parameters ϵ_{ii} and σ_{ii} are determined for individual atom types and then combined via the rules[1] shown in equations 2.12 and 2.13 to obtain the parameters used in the potential function [52, 135, 137, 138].

$$\sigma_{ij} = \frac{\sigma_{ii} + \sigma_{jj}}{2} \tag{2.12}$$

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}} \tag{2.13}$$

The correct choice of the parameters involved in equations 2.10 and 2.11 allows to accurately reproduce all the complex panorama of non-bonded interactions, including also hydrogen bonds and hydrophobic interactions, although they are not taken into account explicitly [135, 138].

2.3 MD algorithms

The idea behind MD simulations is to study the time evolution of a physical system by integrating Newton's motion equation for each particle contained in the system. In this scheme, an approximate potential is used to mimic the interactions in the system and the integration is performed with an appropriate algorithm. The equation of motion for a system of N particles can be written as:

$$m\ddot{\mathbf{r}}_i(t) = \mathbf{F}_i(t), \quad i = 1, ..., N \tag{2.14}$$

[1] Different combining rules exist for the LJ potential. Here we refer to the Lorentz-Berthelot rules, which are the simplest and the most used in this context [135, 137, 138].

where, at time t, the force $\mathbf{F}_i(t)$ acting on particle i depends on the position of the other N-1 particles. This can be explicitly shown writing equation 2.4 in a different form:

$$\mathbf{F}_i = -\nabla_i \sum_{i=1}^{N} \sum_{j>i}^{N} U(r_{ij}(t)) \tag{2.15}$$

with $r_{ij}(t) = |\mathbf{r}_i(t) - \mathbf{r}_j(t)|$. So, after specifying the initial conditions for positions and velocities of each particle, the trajectory can be easily calculated. The force on each particle can be computed using equation 2.15 and the position of the particle can be calculated from equation 2.14.
However, computing the position $\mathbf{r}_i(t)$ requires the integration of equation 2.14 and several fast and robust algorithms have been developed for this scope [55, 56, 132, 139]. Here only the simplest of them is presented, known as the Verlet algorithm [131, 140, 141], which derives from the Taylor expansion about time t of the coordinate $\mathbf{r}_i(t)$ of a particle:

$$\mathbf{r}_i(t + \Delta T) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{\mathbf{F}_i(t)}{2m}\Delta t^2 + \frac{\dddot{\mathbf{r}}_i}{3!}\Delta t^3 + O(\Delta t^4). \tag{2.16}$$

Similarly,

$$\mathbf{r}_i(t - \Delta T) = \mathbf{r}_i(t) - \mathbf{v}_i(t)\Delta t + \frac{\mathbf{F}_i(t)}{2m}\Delta t^2 - \frac{\dddot{\mathbf{r}}_i}{3!}\Delta t^3 + O(\Delta t^4). \tag{2.17}$$

Adding these equations gives:

$$\mathbf{r}_i(t + \Delta T) \approx 2\mathbf{r}_i(t) - \mathbf{r}_i(t - \Delta T) + \frac{\mathbf{F}_i(t)}{2m}\Delta t^2 \tag{2.18}$$

Which leads to an error on position of the order Δt^4. This integration procedure is repeated iteratively each ΔT and the system is evolved in time. In this context the choice of the time step is crucial and it depends on the system in study. A large Δt will result in an unrealistic evolution of the system, causing instabilities in the simulation, while a time step which is too small can result in a "waste" of computational resources to perform a too long simulation to observe meaningful events. For classical MD simulations an appropriate time step is of the order or 1-2 *fs* [142, 143], although more sophisticated strategies exist to allow the use of larger time steps [144].

Although equation 2.18 can be used to compute the trajectory of a set of particles, it has a major drawback: it doesn't explicitly compute the velocities of the particles. In fact, explicitly computing velocities, although not necessary to calculate the trajectory, allows to calculate the total kinetic energy of the system. Over the years, several variations of the original Verlet algorithm have been proposed in order to introduce an explicit evaluation of the velocity of particles. Examples include the so called "velocity Verlet" algorithm [131, 140, 145] and the "Leapfrog" method [131, 140, 146].

2.3.1 The ergodic hypothesis

In classical statistical mechanics, a macroscopic system is represented as an ensemble of N particles, interacting according to the laws of classical mechanics. An instantaneous state of the system, called microstate, is represented by a point in the

phase-space as follows:

$$(r(t)^N, q(t)^N) = (\overrightarrow{r(t)}_1, \overrightarrow{r(t)}_2, ..., \overrightarrow{r(t)}_N; \overrightarrow{q(t)}_1, \overrightarrow{q(t)}_2, ..., \overrightarrow{q(t)}_N) \equiv \Gamma(t) \tag{2.19}$$

The evolution of the system is then described by the trajectory in the phase-space $\Gamma(t)_A \rightarrow \Gamma(t)_B$. However, when dealing with a real system, we are usually interested in measuring macroscopic properties of the system, namely the *observables* of the system. It's important to note that macroscopic time scales are larger than microscopic ones. It means that during a macroscopic measurement of the observable O, the system passes through several different microstates. The link between the observed value of O and its values explored during the observation time is represented by equation 2.20, where the observed value is calculated as the time average of the observable on long times related to the microscopic time scales of the system [131, 132, 135]:

$$\overline{O}(t_0, T) = \frac{1}{T} \int_{t_0}^{t_0+T} O(\Gamma(t))dt \tag{2.20}$$

However, the computation of the integral in the above equation would require a detailed knowledge of the microscopic state of the system at time t_0 and of its trajectory during its evolution for the time T. Those requirements are hardly satisfied for large systems and, if the time evolution of a system strongly dependent on its initial state, it would be very hard to make statistical previsions on its evolution. The *ergodic hypothesis* allows to overcome this obstacle. It basically postulates that every hyper-surface (on the phase space) of fixed energy E is completely accessible to every particle having the right energy. This means that if we observe a system for a sufficiently long time T (e.g. during a MD simulation), the time average in equation 2.20 will depend only on the energy of the system and it will assume the same value for every trajectory of energy E, independently from the microscopic state of the system at t_0. So, for an *ergodic system*, we can state that:

$$\overline{O} = \lim_{T \rightarrow +\infty} \frac{1}{T} \int_{t_0}^{t_0+T} O(\Gamma(t))dt = \int O(\Gamma)\rho_{mc}(\Gamma)d\Gamma = \langle O \rangle \tag{2.21}$$

where $\rho_{mc}(\Gamma)$ is the probability density of each microstate of energy E [131]. This hypothesis has two important advantages for MD simulations [147]:

- The statistically averaged properties of the system in study are accessible through MD simulations, that are aimed to generate trajectories;
- If the simulation is long enough, the time averaged properties become independent from the initial conditions of the system.

Although it is a plausible hypothesis, which is assumed to be valid for the majority of biological systems, there are cases in which this hypothesis is not satisfied [131, 132].

2.3.2 Periodic boundary conditions

Usually, MD simulations are performed to study the properties of a system in bulk, or, more formally, in the thermodynamic limit $N \rightarrow \infty$. However, only systems with a finite and usually relatively small number of atoms (normally less than 10^6, although this limit has recently been pushed up to a billions of atoms [148]) can be simulated, in order to contain the computational effort of the simulation. This

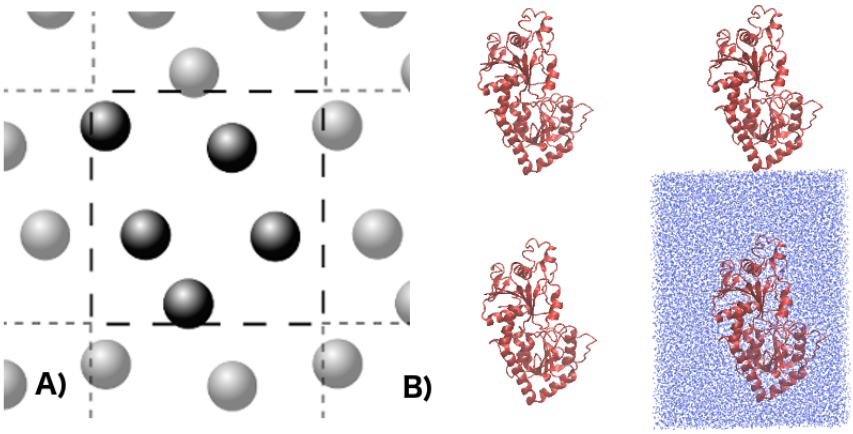

Figure 2.3: PBC: Figure A) shows a 2D lattice, highlighting the particles in the original box and their images in the surrounding boxes. Figure B) is a 2D representation of periodic simulation boxes of a protein. The aqueous environment is showed for one of the boxes.

raises the problem of surface effects, which are substantially due the fact that surface molecules interact with less molecules than bulk ones. To ensure that edge effects will affect to a minor extent the bulk properties of the system, different strategies have been developed. One of the most used is simulating the system under periodic boundary conditions (PBC) [131].
The idea is to simulate the original (finite) simulation box surrounded by a series copies of it, arranged in a three-dimensional infinite lattice, as shown in figure 2.3.

Particles of the surrounding boxes, refereed as *images*, moves in the exactly same way of those in the original box. In this way, when, during a simulation, a particle moves out from the original box, a particle from one of its images enters the box on the opposite side, assuring that the total number of particles in the original box is conserved. When performing a simulation under PBC, however, long-ranged interactions between the system in the central box and its periodic images should also be taken into account. VDW interactions decay very rapidly and so they can be cut at a given cutoff (typically 9-12 Å) and the regions beyond the cutoff treated as homogeneous medium by employing averaged LJ parameters [131]. Electrostatic interactions, on the contrary, are long-ranged and the usage of a cutoff is very likely to introduce serious errors in the force calculation. Instead, it is important to include the force contribution from all the particles involved: the ones in the original simulation box as well as their images in the surrounding boxes [131, 132, 135]. One of the most used method to deal with this issue is the so-called *Ewald summation* [149], which will be addressed in the following section.

2.3.2.1 The Ewald summation

Let's consider a system of N interacting particles under periodic boundary conditions in a cubic box of diameter L ($V = L^3$) and suppose we are interested to evalutate the

total potential energy of the system. Under PBC conditions equation 2.10 becomes:

$$U^{electrostatic} = \frac{1}{2} \sum_{i,j=1}^{N} \sum_{m \in \mathbb{Z}} u(r_{ij} + mL) \tag{2.22}$$

where $u(r_{ij} + mL)$ indicates the potential energy on the particle i due to the electrostatic interaction with the other particles [131, 132, 135]. The radius $r_{ij} = \mathbf{r}_i - r_j$ indicates the distance between the two interacting particles (the case with $i = j$ should be omitted) and m is an index used to consider the periodic images of the particle of the original box (for which $m = 0$).

Writing explicitly the form of $u(r_{ij} + mL)$ in the previous equation leads to:

$$U^{electrostatic} = \frac{1}{2} \sum_{i,j=1}^{N} \sum_{m \in \mathbb{Z}} \frac{q_i q_j}{|r_{ij} + mL|} \tag{2.23}$$

Equation 2.23 represents a well-defined electrostatic problem, however, such computation is not trivial. The first problem in the summation above is that the series is not absolutely convergent, meaning that it is not well-defined unless we specific the way in which the calculation is performed (i.e. the order in which we sum up the terms). A natural choice could be to take the simulation boxes as (roughly) spherical layers but this choice leads to an extremely slow convergence which is not desirable. The second problem is that eq. 2.23 is a sum over $N(N-1)/2$ terms, which, in the case of biological systems, results in an enormous computational cost to perform that calculation, scaling as $O(N^2)$.

To overcome this limitation, the summation in the equation 2.23 can be split into two components, using the identity:

$$\frac{1}{x} = \frac{f(x)}{x} + \frac{1 - f(x)}{x} \tag{2.24}$$

with $x = |r_{ij} + mL|$.

In this way, the initial slowly conditionally convergent series has been converted into two quickly convergent terms: a term which quickly converges in the real space and a term which can be accurately treated in Fourier space.

The idea behind this method can be explained as follows: $f(x)$, which is a generic function of the coordinate x, can be chosen so that $\frac{f(x)}{x}$ is negligible beyond a given (small) cutoff r_{max} while $\frac{1-f(x)}{x}$ will be a slowly varying function of x, which means that its Fourier transform can be well represented by only a few reciprocal vectors [131, 132, 135]. The usual choice for $f(x)$ is the complementary error function erf(x) $= (2/\sqrt{\pi}) \int_x^\infty \exp(-t^2) dt$.

This method not only resolves the convergence problem but it also reduces the computational cost of eq. 2.23, which, in this way, scales as $O(N^{3/2})$. For large systems, however, this approach is still expensive and more sophisticated methods are available to deal with long-range interactions. An example is represented by the Particle-Mesh Ewald method [150] in which the computational cost scales as $O(N \log N)$.

2.4 Enhanced sampling simulations

As discussed above, classical MD simulations applied to biological systems are a powerful tool to explore the microscopic behaviour of complex systems and important successes have been achieved in the last decades. However, this technique may not be suitable to study phenomena which involve transitions in a complex free energy landscape (FEL)[2]. If a system displays a FEL featuring a series of disconnected basins with a high occupancy probability, separated by regions in which the probability of occupancy is very low, the system is defined *metastable* [80, 152–154]. Meta-stable systems spend the majority of time in the disconnected regions with high occupancy probability (i.e. in a sharp minimum, but not necessarily the lowest energy one), with slow-rate transitions allowing the system to explore other regions. Transitions of this kind are collectively named *rare events*. Examples of biological systems featuring rare transitions include protein-folding, gating mechanisms for ionic channels, protein-ligand interactions, etc.
In our case of interest, namely protein-ligand interactions, MD limitations are manly related to the sampling of conformational changes of receptor's structure since some conformational states can be separated by energetic barriers much greater than the energy associated to thermal fluctuations.

To overcome this limitation, several strategies have been developed to accelerate the observations of these rare events by forcing proteins to explore larger portions of its conformational space, discouraging the sampling of already visited regions (and thus helping the system to escape from energy basins in which it was eventually trapped) [66, 155, 156]. The choice of the strategy to use strictly depends on the specific problem to address and on our knowledge of the system [59, 74].

A large class of such methods works by adding a fictitious potential to the real (free) energy landscape of the system, reducing the depth of its minima. In this way, the energy landscape felt by the system is flattered compared to the real one, allowing thermal fluctuations to easily overcome smaller energy barrier of basins in which the system gets stuck. Examples of these strategies include *accelerated molecular dynamics* (aMD) [62] and *metadynamics* [80]. Other methods, such as simulated annealing [157, 158], work by raising the temperature of the system so that the increased thermal energy will (more) easily trigger the transitions. In this panorama, a widespread method is the *temperature replica-exchange MD* (T-REMD), in which a set of MD simulations are run at different temperatures and at predefined time intervals, conformations sampled in neighbouring pairs of replicas are exchanged [159–161]. Still in the context of REMD, approaches based on different schemes are also possible, such as the Hamiltonian REMD (H-REMD), in which the replicas are run with different force-fields [162–164].
In the following, however, we'll focus on metadynamics, which is the strategy used in the works presented here.

[2]The free energy surface can be simply thought as a multi-dimensional free energy profile of a system. See refs. [129, 147, 151] for details.

2.4.1 Dimensional reduction

Let us consider a system of particles of coordinates x in the space Ω, where x can be the usual cartesian coordinates $\vec{r}$ or some other coordinate, coupled with a thermostat bath at temperature T. The system evolves according to the laws of thermodynamics following the canonical equilibrium distribution [131, 147, 165]:

$$P(x) = \frac{1}{Z} e^{-\beta V(x)} \tag{2.25}$$

Where $\beta = \frac{1}{K_B T}$ with K_B the Boltzmann constant and $Z = \int dx e^{-\beta V(x)}$ the partition function of the system. Biological systems are characterized by a large number of atoms, meaning that $P(x)$ has a huge dimensionality. To overcome this problem, what can be done is to consider some reaction coordinate, namely some *collective variable* (CV) instead of the coordinate x [152]. The idea is to study the system through some "collective coordinates" which provide a coarse-grained characterization of the system. In this way, if the proper CVs are chosen (i.e. able to describe the transition(s) under investigation), the evolution of the system can be well described using few suitable parameters instead of the huge number of coordinates x, reducing the computational cost of the simulation. Thus, instead of monitoring the full trajectory $x = x(t)$ of the system, only the reduced trajectory $s(t) = s(x(t))$ is analyzed. Instead of $P(x)$, the probability $P(s)$ can be computed:

$$P(s) = \frac{1}{Z} \int dx e^{-\beta V(x)} \delta(s - s(x)) \tag{2.26}$$

For an infinitely long trajectory, equation 2.26 can be evaluted by the histogram of s [166]:

$$P(s) = \lim_{t \to \infty} \frac{1}{t} \int dt \delta(s - s(t)) \tag{2.27}$$

which, for real applications, becomes:

$$P(s) \approx \frac{1}{n \Delta s} \sum_{t=1}^{n} \chi_s(s(t)) \tag{2.28}$$

where $\chi_s(x) = 1$ if x $\in [s, s+\Delta s]$ and zero otherwise, n is the number of histograms and Δs is the width of each of them. If the system is ergodic and it is in equilibrium at temperature T, we can define the Helmholtz free energy of the system in terms of the collective variable s:

$$F(s) = -\frac{1}{\beta} \ln(P(s)) \tag{2.29}$$

Equation 2.29 shows that it is possible to enhance the sampling of rare events by acting of $F(s)$ via $P(s)$. Moreover, since a strong link exists between $P(s)$ and the FEL (eq. 2.31), a straightforward way to enhance the sampling is to bias the plain dynamics of the system by flattening its FEL. This is typically done by a properly chosen bias potential $V_B(s(x))$, depending on the coordinate x via the collective variable $s(x)$. So a fictitious potential term $V_B(s(x))$ can be added to the potential $V(x)$ allowing the system to "jump" between states separated by higher energetic

barriers. A biased probability distribution can be defined in the same fashion of eq. 2.25:

$$P_B(x) = \frac{1}{Z_B} e^{-\beta(V(x)+V_B(s))} \tag{2.30}$$

where Z_B is the canonical partition function for the new potential $V(x) + V_B(s)$. With the same approach used for eq. 2.26, $P_B(s)$ can be computed and then related to $P(s)$ as follows [80, 131, 154, 165]:

$$P_B(s) = \frac{1}{Z_B} \int dx e^{-\beta(V(x)+V_B(s))} \delta(s - s(x)) = \tag{2.31}$$

$$= \frac{Z}{Z_B} e^{-\beta V_B(s)} \frac{1}{Z} \int dx e^{-\beta V(x)} \delta(s - s(x)) = \tag{2.32}$$

$$= \frac{Z}{Z_B} e^{-\beta V_B(s)} P(s) \tag{2.33}$$

In this way, $P(s)$ can be evalutated as:

$$P(s) = \frac{Z_B}{Z} e^{\beta V_B(s)} P_B(s) \tag{2.34}$$

Finally, using eq. 2.29, the free energy $F(s)$ can now be calculated with respect to the biased probability $P_B(s)$. In order to accelerate the sampling of rare events, the idea would be to find a suitable $V_B(s)$ so that the free energy profile becomes flat [154]. Indeed, the height of the maxima of $F(s)$ is related to the rate of interconversions between the different states of the system. Higher maxima (i.e. higher barriers) are associated to a lower interconversion rate, while lower barriers result in faster interconversions. If a proper bias potential is found and the energy profile becomes ideally flat, the so-called rare events will become accessible also to MD simulations of finite time. It can be shown that a flat free energy profile can be obtained choosing $V_B(s) = -F(s)$, allowing, diffusive transitions in a barrierless conformational space. However, for real systems $F(s)$ is rarerly known a priori, so the main problem is how to construct $V_B(s)$ without a detailed knowledge of the energetic profile of the system.

2.4.2 Metadynamics

Metadynamics (META) [152, 154], is a computational technique which aims to enhance the sampling of rare events and to yield, at least in principle, to the exact free energy profile along the trajectory of one or more CVs.

META acts by flattening the effective free energy of the system during the simulation, preventing the system to be trapped in local minima and being able to explore the entire free energy surface, leading, at least in principle, to a diffusive behaviour of the system in the CV(s) space. Over the years, different groups applied META approaches to unveil details of (macro)molecules dynamics in a variety of different problems [152, 154, 167, 168].

Conceptually, the idea is to enhance the sampling by adding to the potential of the system a bias potential $V_B(s)$ acting on a small number of parameters, namely the collective variables. In the case of metadynamics, the bias potential has the form

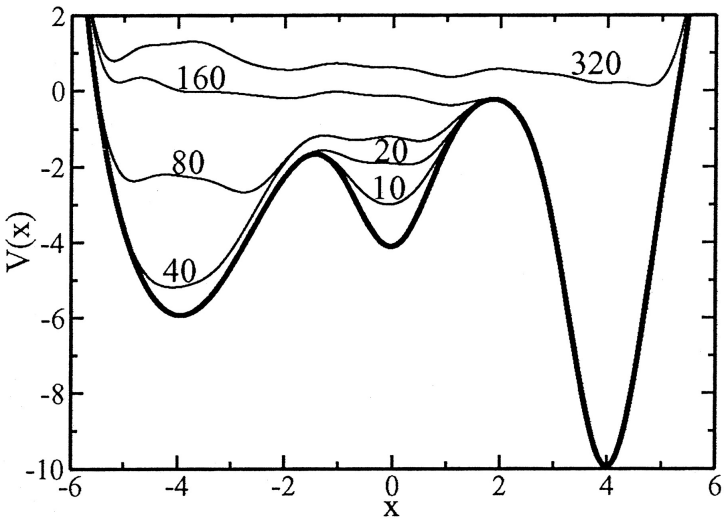

FIGURE 2.4: Schematic representation of the progressive free energy profile filling in a typical META simulation. At $t = 320\tau$ the profile has been completely overcompensated and the system is allowed to make diffusive transitions between the states. Image adapted from ref. [80].

of a history-depend potential built up as a sum of Gaussian distributions centered along the trajectory of the CVs. For a single CV, the bias potential takes the form:

$$V_B(s(x), t) = w \sum_{t'=\tau,2\tau,\ldots}^{t} \exp\left(-\frac{(s(x) - s(x(t')))^2}{2\sigma_s^2}\right) \tag{2.35}$$

where w is the height of the Gaussian distributions, τ^{-1} is the frequency at which the distributions are added, σ_s is their width and t is the total simulation time. In metastable systems, the probability that a system is in a minimum is higher than elsewhere and so, during the simulations, several Gaussian hills are added around the basin in which the system is stuck. In so doing, the hills will increasingly fill the minimum allowing the system to "jump", due to thermal fluctuations, into a close local minimum. Here, the accumulation of hills starts again. In this way, in a sufficiently long simulation, the bias potential compensates all the free energy profile of the system. The novel idea of this method is to use a history-depend potential which keeps memory of all the positions in which the hills have been deposited in order to reconstruct a negative image of the "underlying" free energy (Figure 2.4). More precisely, the important assumption made in this case is that the history-depend bias potential made up of the Gaussians deposited up to the time t is an unbiased estimator for the free energy in the regions explored. It means that, after a transient time t_{eq} needed for the potential to fill all the free energy minima of the system, the bias potential $V_B(s, t)$ will show deviations from $-F(s)$ which will become increasingly smaller as t increases, so that $V(s, t \to \infty) = -F(s) + C$, where C is an irrelevant additive constant (which increases with time) [152, 154, 169].

However, although empirically verified for several systems and also formally proved [64], in actual facts the previous assumption remains an approximation. Since a completely flat free energy profile is very unlikely to be achieved using gaussians of fixed

height, hills are continuously added and the bias potential doesn't exactly converge to $-F(s)$, but it rather oscillates around it. In particular, the accuracy of the free energy (approximative) reconstruction depends on the choice of the three parameters discussed above, w, τ and σ_s [170]. If the Gaussians are large, the exploration of the free energy surface will be fast but at the same time the reconstructed profile will be affected by large errors. On the contrary, if the Gaussians are sharply peaked (small width) or infrequently deposited, the simulation will take a longer time but the reconstruction will be more accurate [170]. Typically, to choose the width of the hills, an unbiased preliminary MD simulation is performed and σ_s is chosen to be of the same order of the standard deviation of the CV in that simulation, in order for the hills to be as large as a fraction of the bottom of the local minima in which the system is trapped. More precisely, in refs. [154, 168] it has been shown that the error on the reconstructed free energy profile depends on the ratio $\omega = w/\tau$, named deposition stride, and not on w and τ separately. Moreover, another major disadvantage of standard metadynamics is that the computational cost increases exponentially with the number of CVs and it has been proven using a large number of CVs has negative consequences on the systematic error as well [152, 154]. So, as a rule of thumb, no more than 3 CVs can be used together in a simulation. On the other hand, neglecting a relevant CV in the simulation can lead to instabilities in the system and in large errors in the reconstruction of the free energy profile.
To overcome the two main limitations described above, two different different approaches have been developed: the *well-tempered metadynamics* [66, 171] and the *bias-exchange metadynamics* [66, 169, 172]. They will be briefly overviewed in the following sections.

2.4.2.1 Well-tempered metadynamics

Well-tempered metadynamics (WT-META) provides a solution to the convergence problem. In this scheme, the bias deposition rate decreases with time. In order to achieve this, a different expression for the bias potential is used compared to standard metadynamics:

$$V_B(s,t) = k_B \Delta T \ln\left(\frac{1+\omega N(s,t)}{k_B \Delta T}\right) \tag{2.36}$$

where ΔT is an input parameter with the dimension of a temperature and $N(s,t)$ is the histogram of the collective variable s collected during the simulation, namely $N(s,t) = \int_0^t \delta_{s,s(t')} dt'$.
The time derivative of $V_B(s,t)$ reads:

$$\dot{V}_B(s,t) = \frac{\omega \delta_{s,s(t)}}{1+\frac{\omega N(s,t)}{k_B \Delta T}} = \omega e^{-\frac{V_B(s,t)}{k_B \Delta T}} \delta_{s,s(t)} \tag{2.37}$$

So, the addition of new hills is exponentially switched off by the growth of the bias potential, as shown in eq. 2.37. This new strategy can be implemented in the standard metadynamics by rescaling the height of the Gaussians deposited so that it is decreased as time increases. In standard metadynamics we saw that the height of the hills was a constant, while, in this case, it becomes time dependent as follows:

$$W(t) = w\tau \exp\left(-\frac{V_B(s,t)}{k_B \Delta T}\right) \tag{2.38}$$

In this way, $\frac{\partial V_B(s,t)}{\partial t} \to 0$ with $t \to \infty$.

The different choice for the bias potential gives to well-tempered metadynamics two important key features with respect to standard metadynamics [152, 154]:

- The deposition rate decreases as $1/t$ and the bias potential converges to its limiting value in a single run:

  $$V_B(s, t \to \infty) = -\frac{\Delta T}{T + \Delta T} F(S) + C$$

  with C an immaterial constant. The factor $\gamma = \frac{T+\Delta T}{T}$ is often refereed as *bias factor*.

- In the long time limit, the probability distribution of the CVs becomes

  $$P(s) \propto e^{-\frac{F(s)}{k_B(T+\Delta T)}}$$

  allowing to control the extent of the FES exploration by tuning the bias factor. For $\Delta T \to 0$ the system behaves as in ordinary MD simulations, while the limit $\Delta T \to \infty$ corresponds to the standard metadynamics, since the scaling factor vanishes.

2.4.2.2 Bias-exchange metadynamics

Bias-exchange metadynamics (BE-META) is based on the combination of the standard metadynamics and replica exchange method and it allows to use a large set of CVs, overcoming the limits of standard metadynamics. The idea is to perform multiple standard metadynamics simulations (called *walkers*) in parallel and at the same temperature, where each replica is biased with a history-depend potential acting on one or more CVs. The sampling is enhanced by attempting, at fixed time intervals, exchanges between the bias potentials of two different replicas. The exchanges are then accepted or rejected according to the Metropolis criterion, with a probability [154, 169, 172]

$$min\left\{1, \exp\left[\frac{1}{k_B T}(V_B^a(x^a,t) + V_B^b(x^b,t) - V_B^a(x^b,t) - V_B^b(x^a,t))\right]\right\} \tag{2.39}$$

where x^a and x^b are the coordinates of the two walkers and $V_B^{a(b)}$ is the potential energy of the system $a(b)$. If the move is accepted, the trajectory that was biased on one of the two coordinates, continues its evolution biased on the other one. The great improvement of this strategy compared to the standard approach is that it allows each run to benefit of the enhancing power of all the CVs used in the different runs, without affecting significantly the speed of sampling since in this case the computational cost increases linearly with the number of CVs and not exponentially, as in the standard approach [66, 172]. However, a major drawback of this method is that the result of a simulation is not a free energy (hyper)-surface in several dimensions but several, less informative, projections of the free energy surface along each of the CVs [66, 152, 168, 172]

2.5 Molecular docking

As already discussed, the success of docking calculations depends on both the search algorithm and the scoring step. In the following, both aspects will be briefly discussed. Specifically, regarding the searching step, the different methods to treat receptor's flexibility will be addressed in some details.

2.5.1 Partner's flexibility in docking algorithms

The searching step determines the number and the quality of the poses generated in a docking run [2, 69, 72]. It is during this process, indeed, that ligand and receptor structures associate to form a complex. As protein-ligand association involves a huge number of degrees of freedom (DoFs), this is a very delicate step, since an inadeguate treatment of these DoFs will likely lead to the inability of the algorithm to retrieve native-like ligand poses. In general terms, the DoFs involved in protein-ligand association can be categorised as follows: (i) the roto-translation between two rigid molecules, involving six degrees of freedom for each molecule; (ii) the conformational degrees of freedom of both partners, reflecting their geometrical fluctuations during the binding process and thus related to their intrinsic and induced flexibilities [2, 15, 59, 173, 174]; (iii) the position and displacement of solvent molecules, that have been proved to play often a crucial role in determining the correct protein-ligand geometry [175–177] also and contributing to the stability of the pose. According to how the conformational changes are treated, docking strategies are typically classified into three classes [67, 71, 178–180]: (i) rigid-body docking, that according to the lock-and-key model, considers both binding partners as rigid bodies, and thus includes only the roto-traslational DoFs in search step. Clearly, this strategy typically works only for cases in which the bound conformations of both the partners are available; (ii) semi-flexible docking, in which one of the partners is kept rigid and only the flexibility of the other one is considered. For example, if the true bound ligand conformation is known, this strategy is suitable to account for receptor plasticity; (iii) full flexible docking, in which the flexibility of both partners is considered. However, also in the case of full flexible docking, an exhaustive treatment of all DoFs is out of reach even for most advanced and dedicated settings, and heuristic strategies are generally used to reduce the dimensionality of the problem.

While the inclusion of small ligands flexibility in molecular docking is nowadays relatively straightforward and encoded within several software packages, accounting for protein plasticity has proven to be a much more challenging task, particularly when rearrangements involving subtle, large, or even secondary structure conformational changes accompany the formation of the complex (Figure 1.1). In the following, we give an overview of some of the most commonly used techniques to account for receptor flexibility in molecular docking, adopting the classification proposed by Antunes and coworkers in ref. [15], and represented in Figure 2.5.

- **Soft docking**

  The idea behind soft docking (Figure 1.1) is to account for protein plasticity in an implicit way, allowing for small ligand-receptor steric clashes during the search for possible binding poses [54, 77, 174]. This is typically achieved

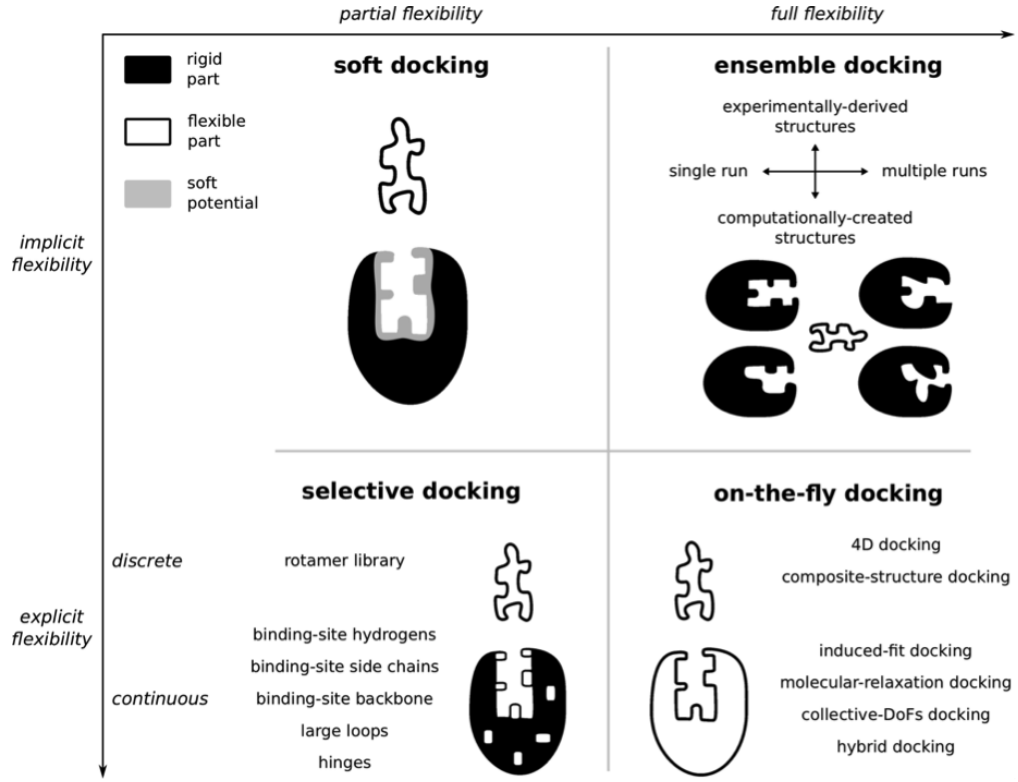

FIGURE 2.5: Classification of different approaches to include receptor flexibility in docking algorithms. From top to bottom: approaches including flexibility in an implicit/explicit way. From left to right: approaches including partial/full flexibility. Image taken from ref. [15].

by "softening" the van der Waals (vdW) potentials of a (sub)set of receptor atoms in its putative binding region. VdW potentials in docking algorithms are usually modelled through a Lennard-Jones semi-empirical function ($V(r)$) increasing very rapidly at short inter-atomic distances. The functional form of $V(r)$ is shown in eq. 2.40, where A and σ are known parameters for each pair of interacting atoms, r represents their interatomic distance and m and n determine the slope of the potential. Standard values of the latter parameters are 12 and 6, respectively [181]. For this reason such a potential is also named the "12-6 law". Typical soft-docking approaches consist in introducing a smoother potential with respect to the standard 12-6 LJ potential. For this reason the values of m and n are usually set to smaller numbers, such as 9-6 as done by Ferrari et al. in ref. [76], or 8-4 as implemented for example in the GOLD software [54, 182]. This reduces the penalties associated to steric clashes allowing for small (≈ 1 Å) ligand-receptor overlaps.

$$V(r) = A\left[\left(\frac{\sigma}{r}\right)^m - \left(\frac{\sigma}{r}\right)^n\right] \tag{2.40}$$

As such, despite not adding any computational cost with respect to rigid docking, it is typically used only to account for small conformational changes, such as of small side-chains reorientations or very limited backbone motions, while remains

hardly able to reproduce major conformational changes involving larger protein backbone motions and/or extended secondary structure rearrangements [174, 183]. For this reason, when adopted for cases in which the details of the binding event are not known a priori, it is usually encoded as preliminary step within more complex approaches [184, 185]. Finally, since the favoured ligand/receptor overlaps may create structural artefacts in the generated complexes, several reports shown that this approach can increase the number of false positives with respect to standard rigid docking [15].

- **Selective Docking**

  Selective docking (Figure 1.1) is another strategy to include target flexibility in docking calculations in a computationally treatable way, by explicitly exploring only a few critical DoFs of the receptor [186]. In a typical selective docking run, the location of the BS in known in advance, a common choice is to limit the accounted DoFs only to rotations of dihedral (torsional) angles of selected side chains receptor's putative site. In earlier implementations, these DoFs were explored through a so-called "discrete" approach, using specific rotamer libraries containing a set of predetermined values for the allowed dihedral angles [15, 180, 187, 188]. Those libraries were usually built on the analysis of experimental data so only the most frequent (low-energy) side-chains conformations were included. For this reason, libraries appear well suited to identify binding poses of compounds congeneric or having a similar binding modes to the ones for which structural experimental information is available. On the other hand, the use of rotamer libraries will also inherently bias the results in favour of binding modes compatible with the rotations encoded in the dataset, affecting the search for new or rare ones [45]. To cope with this issue, some authors used a discrete and uniform set of angle values, equally space and not weighted by the statistical occurrence in known structures [189–191]. To reduce the computational cost of this procedure the search (sub)space is also usually restricted according with geometrical constraints imposed by the structure of the BS. Current implementations of selective docking allow the user to select the DoFs to be explicitly (and in a virtually continuous way) accounted for during the calculation. It is also typical to include (small) backbone rearrangements of the whole protein or of specific regions [192] or to explicitly treat hydrogen orientations as it is well-known that H-bonds play a crucial role in binding events [15].

- **On-the-fly docking**

  At odd with the two previously addressed methods, this class of methods aims to explicitly account for the (virtually) full receptor flexibility during docking (Figure 1.1). Due to the extremely high dimensionality of the problem, heuristic strategies are used in order to limit the associated computational cost. Over the years, a huge number of different on-the-fly approaches have been developed; in the following we describe only two among the most frequently adopted strategies, referring the interested reader to more exhaustive works on the subject [15, 59, 193, 194].

Induced-fit docking. Built according to the induced fit mechanism of MR [15, 184], the induced-fit docking typically involves [2, 15, 173, 184, 195]: (i) a first step in which soft-docking calculations are performed to place the ligand into receptor's BS; (ii) a second step of structural optimisation, usually involving reorientations of selected side-chains by means of rotamer libraries or by means of Molecular Mechanics [196], Monte Carlo [58], or Molecular Dynamics [128] calculations. Induced fit docking has been successfully used in several studies, becoming a very powerful tool in the field of ligand-receptor association [15, 173]. Specific recipes developed by different groups generally differ in the implementation details of these steps and often include several additional intermediate ones (see e.g. refs. [184, 195, 197, 198]).

Molecular-relaxation docking. Following this approach, docking is treated as a molecular relaxation problem. Starting from a certain structure of the ligand-receptor complex, generally obtained by means of rigid/soft-body docking, molecular optimisation [15, 193] can be performed by means of a plethora of different techniques, such as energy minimization [178], Monte Carlo [199] or Molecular Dynamics [59, 60, 200, 201]. For an example of one of such approaches employed in the context of the blind docking challenge D2R Grand Challenge 2, see for instance ref. [202].

- **Ensemble docking**

  At odd with all the other methods discussed so far, in this approach full receptor's flexibility is accounted prior to docking calculations (Figure 1.1). Instead of a single structure, a set of different receptor conformations is given to the docking algorithm. In so doing, this approach follows is based on what predicted by the "conformational selection" model, that ligand bind will occur on conformation already featuring the "right" (i.e. near-holo) conformation. For this reason, receptor is usually treated as rigid during the docking stage of ensemble docking calculations, as its flexibility has been already addressed in a previous step.

  Over the years, a number of works shown that, in absence of the true holo-like receptor conformation with respect to the ligand of interest, ensemble-docking approaches, either with experimentally or computationally determined conformations, improved docking and virtual screening performances with respect to using a single-structure docking calculations [59, 60, 74, 78, 203].

  In the case of experimentally-derived conformations, a common strategy is to use receptor holo structures complexed with ligands similar to the one in study, as this has been proved to increase docking performance with respect of using unbound protein conformations [7]. However, since ligand-induced/recognised rearrangements are extremely ligand-specific, the improvement is not guaranteed when the included receptor structures are obtained from complexes with ligands belonging to different chemotypes than the one of interest [7, 59, 72, 203]. Moreover, compared to the druggable genome [60, 204], accounting to around the 20% of all protein coding genes [205], the experimentally-determined bound complexes are limited and biased toward a small number of studied cases, for example involving pharmaceutically relevant receptors [45], having a dramatic

impact on the chemical diversity of putative lead compounds that can be successfully studied in virtual screening campaigns. Furthermore, it is important to mention that using an ensemble of structures coming from complexes with ligands similar to the one in study, also requires that the scoring function used to rank the generated poses is competent to discriminate between biologically active/inactive binding modes. Unfortunately this is hardly guaranteed as currently used scoring functions are not usually sufficiently reliable to provide accurate ranks for different binding modes associated to slightly different protein/ligand conformers, as it can happen with ensemble docking calculations, where the aim is to have a native-like pose of certain ligand ranked on top but using a pool of receptor conformations coming from complexes with different ligands [7, 59, 60, 74, 203].

On the other hand, computational strategies have been developed to generate protein conformations, such as Monte Carlo [58, 199], Molecular dynamics [60], and Normal mode analysis [185, 206], making ensemble docking by far the most widely used strategy to account for receptor plasticity in SBDD approaches [59, 110, 173, 207].

One of the main recipes used for SBDD calculations is the Relaxed Complex Scheme (RCS) introduced by Lin et al. [208] whereby a multiple-run docking calculation is performed on a pool of receptor conformations of the unbound protein generated by MD simulations. Although the efficiency of this strategy in improving docking performance and VS efforts has been largely demonstrated [59, 74, 174, 201, 209], a central issue is how to determine the optimal number of structure to include in the ensemble [59, 201]. In principle using large ensembles or even the whole generated MD trajectory might appear reasonable, however this has been linked to the severe risk of generating many false positives/negatives [59, 74], due to the (already mentioned) limits of the current SFs to discriminate between biologically relevant native-like poses and inactive ones, as reported and discussed in refs. [2, 69, 72, 73].

For this reason and to reduce the computational cost of the calculations, typical RCS approaches involve the selection of a (small) number of MD snapshots either in a (pseudo)casual way or by means a cluster analysis (CA) [174, 210, 211]. Common cluster-analysis techniques involve the use of the RMSD (Root Mean Square Deviation) metrics evaluated on the putative receptor BS, either coupled with agglomerative [210, 212–215] or divisive [210, 213, 216–219] clustering approaches. However, for successful docking calculations the morphology of the pocket is essential, meaning the volume of the pocket together with the exact location and orientation of all the interacting side-chains of the BS residues. For this reason, as highlighted by others and shown also in this thesis, the usage of RMSD-based cluster analysis is not always the best strategy to capture efficiently the structural diversity sampled during the MD [24, 220]. Other clustering metrics developed include the ones developed by Motta and Bonati [24] and Osguthorpe and coworkers [220] in which the conformations are grouped on the basis of the volume of the putative site and the one addressed in this work in which the clustering is based on the usage of 3D shape descriptors of the binding pocket [22].

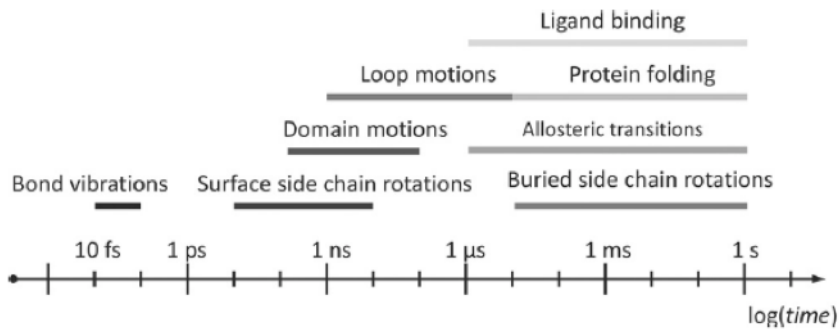

FIGURE 2.6: Typical timescales associated to protein dynamics. Large conformational changes, usually associated to ligand binding events, involve time-scales going from microseconds to seconds. Taken from ref. [59].

Another important issue when dealing with MD-derived ensembles, is how long MD simulations should last to obtain a statistically converged set of conformational states. Already simulations of tenths of nanoseconds has been proven to be effective for some applications [74], although the at least the (micro)second scale is typically needed to observe large conformational changes as ligand-induced/recognised ones (Figure 2.6) [59].

Moreover, for real applications, the extent of conformational changes associated to a specific MR event might not be known in advance. In this scenario, in absence of data on the structure/energetics of the true holo conformation which could drive the sampling, (long) MD simulations [56, 59, 156, 221], approaching at least the (micro)second time scale should be carried out, so to include, in ensemble of structures, also high-energy states featuring large and/or somehow kinetically unfavourable conformational changes with respect to the apo conformation [74, 200, 201, 221]. Including in the ensemble also short-living (high energy) (meta)states is thus a way to mimic what in true biological environment would be induced/stabilised by ligand interaction with the aim to catch also states resembling the true holo-like conformation of the protein under investigation. For this reason, due to the time limitations of plain MD, conformations generated by means of enhanced sampling techniques [155] also often included in the ensembles.

2.5.2 The scoring in docking algorithms

In order to produce trustworthy results, an exhaustive sampling of the conformational space associated to possible complex geometries must be coupled with a reliable ranking strategy [222, 223]. In an ideal scenario, such ranking would not only be a scoring parameter, simply dissecting the most likely to occur poses from the less favoured ones, but it should also accurately estimate the binding affinities (ΔG) associated to the different binding poses. Computational accurate methods exist to perform such task, such as free energy perturbation [224, 225] and thermodynamic integration [226], just to cite two options. Unfortunately, despite being generally recognised as accurate methods, they require a substantial computational cost,

making them virtually unusable in most CADD applications, where many ligand-receptor complexes can involved. For this reason, since earlier docking studies, faster but less accurate ranking methods have been preferred [222]. Among them, the most widespread strategy is based on the concept of ad hoc constructed functions, named "scoring function" (SFs) [73, 222] depending on an handful set of parameters and taking into account the most important interactions to the binding. Several assumptions and approximations are usually made when constructing a scoring function, in a sort of trade-off between making the scoring step computationally feasible but at the same time enough accurate to produce reliable results [46, 73, 75, 223]. Re-scoring strategies however exist, so that binding free energy (or just the ranking) of docking-generated complexes is evaluated after the docking stage, often also following a process of structural refinement. The re-scoring step, in this context, is independent from the docking strategy used in the previous step. It can be thus carried out by means of SFs or with different methods. In the latter case, commonly used strategies are the MM/PBSA and MM/GBSA [227], intermediate for both accuracy and computational cost between SFs and and strict alchemical methods, although recent reports showed how these methods don't always improve the results (see e.g. ref. [228]).

To evaluate the reliability of SFs, three criteria are considered: (i) Its ability to identify to rank native-like poses on top; (ii) its ability to discriminate between potential binders and inactive ligands from large dataset of compounds typically used in VS efforts [2, 15, 73, 85] and (iii) its ability to estimate binding free energy values matching with the experimental data; No SF meeting all the above-mentioned criteria when tested on large and diverse ligand-receptor datasets has been developed to date. Indeed, despite the large number of SFs proposed over the years, the scoring problem still represents a major challenge in CADD [73, 173, 222]. Among the most important SFs limitations, the dependence of its accuracy on the specificity of binding interactions actually involved in the binding process is a critical one. For instance, a SF not explicitly considering (de)solvation effects might be accurate for cases in which the displacement of structural waters upon binding is minor but it is likely to produce unreliable results for cases where this process plays a crucial role [46, 73, 75, 222, 223]. In this sense, SFs suffer from the difficulties associated to lack of a strong theoretical model allowing to describe, via simple (and easy to computationally treat) functions, phenomena such as (de)solvation effects, entropy contributions and hydrophobic interactions. For more details, including a detailed description of the most commonly used SFs, see for instance refs. [75, 222, 223, 229–233].

2.6 Druggability assessment

In the context of ligand binding, the term *druggability* refers to the ability of a biological macromolecule to accommodate compounds with drug-like properties leading to a modulation of its function [234, 235]. With only about the 20% of the human genome representing druggable targets, and only half of those being directly linked to diseases, methods able to predict the druggability of novel targets have become of great help in the early phases of drug discovery [205]. In this work, we estimated the druggability of MD generated conformations of protein using open-source pocket detection package f-pocket [236] able to identify and characterize

putative protein binding sites. The algorithm implemented in f-pocket is based on Voronoi tessellation and alpha spheres and has proven to be stable, fast, and accurate, performing very well on state-of-the-art data sets [236]. Moreover, for each putative binding site identified, f-pocket also estimates its druggability through the calculation of a druggability score, D [237] ranging from 0 for no druggable pocket to 1 for pockets with a high probability to be druggable. To evaluate the druggability of the known binding sites on the proteins investigated in this work, we recorded only D values associated with pockets whose centers of mass were found to be within 6 Å of that of the binding site identified in the experimental structures.

2.7 Analysis methods

2.7.1 RMSD analysis

The root mean square deviation (*RMSD*) is by far the most commonly used quantitative measure of the overall similarity between two (macro)molecular structures [238]. It is defined as the square root of the average of the squared distances between corresponding atoms of x and y and it is used as a measure of the average atomic displacement between two conformations of the same molecule. RMSD values are usually reported in angstrom (Å)[3] and calculated by means of equation 2.41, where the averaging is performed over the n pairs of equivalent atoms selected for the calculations. It is important to note that to use equation 2.41 each structure must be represented as a 3N-length vector of coordinates, where N is the number of atoms of each structure.

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{y}_i\|^2} \tag{2.41}$$

However, before the proper RMSD calculation, the conformations to be compared need to be superposed, as they are often (e.g. when molecular conformations are sampled from molecular dynamics) in different positions and orientations. In order to do so, a preliminary RMSD is calculated between the initial conformations in the original position/orientation and then the optimal roto-traslational motion able to minimize the RMSD is searched. In this way the conformations are superimposed by means of a (set of) rigid motion(s) and placed in the same orientation. To do so, several algorithms exist [239] and have been successfully employed. Examples include the Kabsch algorithm [240, 241] which uses the rotation matrices to represent rotational motions or methods in which such representation is given by means of quaternions [242].

RMSD calculations can be performed for any type and subset of atoms; in this work, a typical choice for the atom types will be using all heavy atoms (i.e. all atoms except hydrogens) while concerning the subset of atoms to consider, calculations will be mostly carried out only for atoms belonging to putative binding site(s) of the proteins.

[3] Note that 1 Å = 10^{-10} m.

2.7.2 Cluster analysis

Another important strategy needed in the framework of EDES approach is the cluster analysis, used to extract a set of representative protein conformations for docking calculations from the MD trajectories. Cluster analysis is a statistical data mining tool by which a set of objects with similar features are grouped together in clusters [210, 211]. The aim of cluster analysis is to thus organise observed data into meaningful structures in order to (i) gain further insight from them; (ii) after having selected the feature(s) of interest, reduce the size of the dataset by creating a new (smaller) one, featuring the same properties of the original one with respect to the feature(s) under study. Key property of a good cluster analysis should be to make points within each cluster similar to each other while points belonging to different clusters dissimilar to the ones in the other clusters. In order to do so, cluster algorithms require the definition of the notion of "similarity", typically a distance measure. Examples of such notions include euclidean, cosine, Jaccardi distances [210, 211].

Concerning MD simulations, the improvement of the computational power, made it possible to simulate large biological systems in the microsecond scale. As a result, modern MD simulations often produce massive amounts of data, typically containing millions of conformations [56]. For this reason, clustering methods have been widely used to group together similar conformational states from MD simulations. In this context, the aim of cluster analysis is to capture a set of conformations representing all the states sampled during the whole simulation. In this way, subsequent studies on those conformations can be focused on a (small) tractable set of structures.

In general, three ingredients are needed to perform a cluster analysis: (i) the clustering algorithm; (ii) the feature(s) to be clustered and the metrics to use, i.e. a proper way to calculate the distances between objects in the feature space; (iii) a handful number of parameters dependent on the specific algorithm chosen (such as the number of clusters, cut-off(s) to be used, etc.).

Over the years, a large variety of clustering strategies have been successfully applied to MD datasets [57, 210, 211, 213, 243–245], elucidating strengths and pitfalls of each approach. Up to date, however no general clustering recipe exists, and the specific clustering procedure to be used is still case-dependent.

One of the most commonly used clustering techniques in the field of MD simulations of biological systems is the so called hierarchical agglomerative clustering (HAC) [210, 212–215], implemented for example in the *cpptraj* [246] and *gromos* modules respectively of the AMBER [136] and GROMACS [247] packages.

Given a MD trajectory of N protein conformations to be clustered, the process of HAC (in its simplest form, as presented by ref. [248]) can be summarized as follows:

1. Given similarity notion and the metrics, calculate the $N \cdot N$ distance (or similarity) matrix. Common choices are the euclidean distance for the first option and the RMSD calculated over the relevant portion of the protein (e.g. the putative binding site(s)) for the latter;

2. Assign each conformation to a cluster (obtaining N clusters), each containing just one item. Define the (dis)similarity notion between two clusters as the one between the items they contain.

3. Find the closest (*most similar*) pair of clusters and merge them into a single cluster, obtaining one cluster less with respect to the ones in the previous step.

4. Compute similarity between the new cluster and each of the previously obtained ones.

5. Iterate steps 3 and 4 until all items are clustered into the desired number of clusters.

In particular, step 3 can be carried out in different ways. Examples of three different typical algorithms used are the single-linkage, complete-linkage and average-linkage approaches, in which the distance between two clusters is defined respectively as the shortest, greatest or average distance among any member of the two clusters. However, more sophisticated approaches exist in which also the statistical properties of clusters are included. Among them, the Ward algorithm [249, 250] is a very popular one. In this case, step 3 is carried out merging the (two) clusters so that the within-cluster variance of the new cluster is minimised.

Another well-known algorithm used for cluster analysis and based on a different philosophy than the HAC is the K-means one [210, 213, 216–219]. K-means is a partitional algorithm in which at the first step all of the objects are in the same cluster[4]. The algorithm then splits the single cluster into smaller ones in a iterative process that stops when each object has been separated into a different cluster or when the requested number of clusters has been achieved.

In particular, referring to the N conformations extracted from an MD trajectory, a typical cluster analysis via K-means involves the following steps:

1. Choose the number of desired clusters (k) and pick up a set of k cluster centers (centroids);

2. Assign each conformation to the closest centroids according to the metrics used. This builds up the first generation of clusters forming a Voronoi partition of the data [251];

3. For each cluster, calculate the new centroids by averaging the feature(s) values used as metrics of its members;

4. Repeat the previous two steps iteratively until the cluster centroids stop changing their positions and become static. Once the clusters become static then k-means clustering algorithm is said to be converged.

However, the algorithm does not guarantee to find the optimum solution, but it rather converges to a local minimum [252]. Indeed it has been shown that the solution strongly depends on the initialisation used. Typical initialisations are based on a random selection of the initial set of centroids from the dataset [253, 254]. This simple approach has the advantage that if we choose points randomly we are more likely to choose a point near a cluster centre since it is where the highest density of points is located. However, there is no guarantee that we will not choose two or more

[4] At odd with HAC, which is said agglomerative and starts in a condition in which all objects are separated into different clusters

points near the centre of the same cluster, thus not maximising the conformational diversity of the generated ensemble of clusters. Indeed, when aiming at that scope, random initialisations is considered one of the most unreliable ones on the basis of a comparison of several alternative algorithms on a range of diverse data sets [255].

In this work, clustering of MD trajectories is performed at the scope of selecting a number of conformations to use for ensemble-docking calculations. Our aim was twofold: (i) maximise the conformational diversity of the ensemble with respect to the shape/volume of the putative binding site(s); (ii) keeping the number of clusters as small as possible.

We found that performing the clustering on the RMSD calculated on a large number of residues (such as the ones lining the BS of a protein) gives a too coarse-grained representation of the conformational motions. Subtle residue reorientations or small changes in the volume of the pocket are hardly accounted with the standard RMSD calculated on the entire BS. In particular we found out that this approach coupled with a small number of conformations retained (compared to the length of the simulation) was not able to capture key BS conformational changes crucial for ligand binding. Indeed, it has been shown that strategies based on a 3D descriptor linked to the pocket shape appear to be more effective in delivering maximally different conformations of the binding site [24, 220].

For this reason we introduced a new clustering strategy not based on the RMSD metrics but on the CVs biased during the metadynamics simulations. As the set of the 4 CVs was proven to be effective in representing the changes of the BS in terms of both the volume and the shape, it was the most natural selection as clustering metrics. Our procedure is based on two different clustering steps: (i) a first step of hierarchical agglomerative clustering; (ii) a second step using the K-means algorithm that is initialised with the centroids being the cluster representatives selected in step (i).

In details, the clustering is performed according to the following procedure (carried out using the R data analysis software [256]):

The distribution of RoG_{BS} values sampled during the MD simulations was binned into 10 equally wide slices, and hierarchical agglomerative clustering (using the built-in function *hclust* of the *cluster* R-package and the Euclidean method to compute the distance matrix) was performed on the selected CVs within each slice, setting the number of generated clusters to $x_i = (N_i/N_{tot}) \cdot N_c$, where N_i, N_{tot}, and N_c are, respectively, the number of structures within the ith slice, the total number of structures in the whole simulation, and the total number of clusters respectively. After that, the clusters for each slice were extracted and used as initialising point for the subsequent K-means step[5]. In this way: (i) the RoG slices serve as a guide for the clustering; (ii) within each slice of conformations with similar RoG, are grouped together on the basis of the volume/shape of the binding site(s), mediated by the CIP CVs; (iii) a K-means step on the whole trajectory serves to group together possible similar structures coming from different RoG slices, so to avoid redundancy and maximise the conformational diversity of the cluster ensemble.

[5] A step by step tutorial on this clustering strategy can be found at the following link: `http://www.bonvinlab.org/education/biomolecular-simulations-2019/Metadynamics_tutorial/` referring to an hands-on tutorial that we gave during the BioExcel Summer School 2019. The files needed to perform the analysis can be found here: `https://github.com/haddocking/EDES`.

2.8 EDES workflow

All the ingredients discussed so far are employed in the method we propose here called *Ensemble Docking with Enhanced sampling of pocket Shape* (*EDES*) [22] exploiting relatively short metadynamics simulations of the apo protein of interest to generate a set of holo-like conformations for ensemble docking. The workflow of the protocol is sketched in Figure 2.7a. First, we identify the putative binding sites on the target proteins. For the purpose of validating the methodology, we identified binding site regions from the structures of the bound complexes. However, we also tested the performance of the EDES method for cases in which the putative pocket was detected by means of site finder web servers, such as COACH-D [257]. Next, we calculate the principal axes of inertia of the binding site. We then use these axes to identify the so called *inertia planes* of the binding site, which are the planes orthogonal to the corresponding inertia axes and passing through the center of mass of the site 2.7b. So, in total three planes are calculated, each identified by two of the inertia axes. Then we perform relatively short bias-exchange, well-tempered metadynamics simulations of the apo protein, using a set of four collective variables (CVs): (i) three (pseudo)contacts across inertia plane (CIP) variables, each defined as the number of contacts between the residues of the binding site on opposite sides of the corresponding inertia plane (fig. 2.7c) and (ii) the gyration radius of the binding site (RoG_{BS}). In particular, CIPs were defined according to the following scheme: for each inertia plane, the residues lining the binding site were split into two lists A and B, according to the positions of the geometrical centers of their backbones on each of the two sides of the plane. Then, the overall number of (pseudo)contacts N_c between the two groups was calculated through a switching function (eq. 2.43) such as the following:

$$N_c = \sum_{i \in A} \sum_{j \in B} s_{ij} \tag{2.42}$$

$$s_{ij} = \left[1 - \left(\frac{r_{ij}}{r_0}\right)^n\right] / \left[1 - \left(\frac{r_{ij}}{r_0}\right)^m\right] \tag{2.43}$$

With $r_0 = 8$ Å, n = 6, and m = 12. We manually ensured that residues from two different groups did not belong to the same secondary structural element. Indeed, this was necessary to avoid the onset of fictitious secondary structure changes leading to very high energy distorted structures. We also use the RoG_{BS} to implement a *windows approach* (fig. 2.7-d) aimed at sampling more effectively and in a controlled manner different shapes of the binding site (possibly mimicking conformational changes induced by ligand binding). Namely, we apply soft walls at RoG_{BS} values that are 7.5% higher and lower than the value measured in the apo X-ray structure (RoG_{BS}^{apo}, corresponding to the center of window 1). Next, from the trajectory of this first window, we randomly select a conformation of the protein whose RoG_{BS} is 5% lower to initiate another MD simulation (corresponding to window 2) with walls centered at ±7.5% RoG_{BS}^{apo} from this new center. We repeat this procedure to generate up to four windows including the first one. This leads to an overall reduction of RoG_{BS} of 15% relative to the center of the first window (RoG_{BS}^{apo}). Despite the arbitrariness of our choice, the performance of EDES is not very sensitive

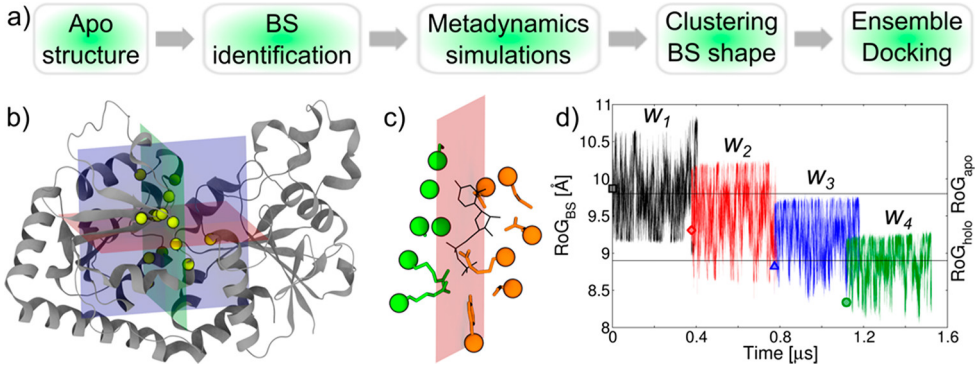

Figure 2.7: Overview of the EDES approach. (a) Workflow of the EDES protocol. (b) Representation of the *inertia planes* (transparent blue, red, and green) calculated at the binding site. The α-carbons of residues lining this site are shown as yellow spheres, and the protein is shown in gray ribbon. (c) Schematic view of the two groups of atoms (orange and green sticks with α-carbons as spheres) considered for the calculation of the number of contacts across one inertia plane; the ligand is also shown in black sticks. (d) Scheme of the "window approach" implemented to enhance in a controlled manner the sampling of conformations associated with different radius of gyration values of the binding site (RoG_{BS}) (the plot refers to simulations of the BGT system). The RoG_{BS} values corresponding to the initial conformation for each window are indicated by a square (w_1), diamond (w_2), triangle (w_3), and circle (w_4). The RoG_{BS} of the apo and holo experimental structures are indicated by horizontal lines. Image from ref. [22].

to the exact choice of three or four windows (and thus to the exact extent of the collapse induced at the binding site, amounting to 10% or 15% of the initial value, respectively). Details of this implementation are better discussed in the next chapter. After the trajectory are generated, we perform an ad-hoc clustering analysis on MD trajectories according to the protocol described in the previous subsection, to select a set of representative protein conformations and finally we use the ensemble of clusters generated to perform ensemble-docking calculations.

In the following, we'll report the performance of EDES in three paradigmatic situations in the panorama of computer-aided drug design. In our first work [22], EDES is tested in a typical re-docking experiment. Namely, we selected three target proteins displaying different extents of conformational changes upon binding and for which previous works showed they represent challenging targets for docking calculations. For each of them, an ensemble of receptor structures is generated according to the original EDES protocol and used for ensemble-docking calculations, where the experimental structures of ligands are re-docked into the EDES generated set of receptor conformations. Docking calculations are carried out by means of two docking packages differing in both the searching and scoring scheme. In our second work, an improved recipe of EDES method is tested in the context of the fourth iteration of the Drug Design Data Resource (D3R) Grand Challenge (GC4). It represents a blind docking challenge where the participants are asked to predict near-native binding poses of a set of 20 ligands against a specific receptor, for which only the primary sequence is given. Moreover, since ligands were made available only by their SMILES code, part of the task was to generate ligand conformers. For this reason we coupled the improved EDES method to a template-based approach to generate ligand conformations. Finally, we report also very preliminary results of a third study, in which we tested an EDES-derived approach to another challenging protein, featuring a very extended binding region and undergoing major conformational changes upon ligand binding. In this case, furthermore, we identified target's binding region only by means of a site-detection software, without exploiting any experimental information of its bound conformation(s).

3.1 EDES in Re-Docking calculations

In the following, we'll assess the sampling performance of the method on three targets that are representative of systems undergoing minor to very large conformational rearrangements upon ligand binding (Figures 3.1 and 3.4). Finally, the set of conformations generated will be used for ensemble-docking calculations.

The first target is the T4 phage β-glucosyltransferase (hereafter BGT) [259] which

undergoes a hinge-bending motion leading to a compaction of its BS upon binding of uridine diphosphate (UDP) compared to the ligand-free structure (Figures 3.1-a,d). This target was chosen because Seeliger and de Groot [23] showed it is a challenging target for re-docking calculations. In their work, they selected 10 targets to assess their workflow based on the generation of a set holo-like conformations of each target by means of an enhanced sampling approach using the tCONCOORD software [260, 261], with the radius of gyration of the holo structure as a bias to drive the sampling. These conformations were then used in ensemble-docking calculations. In eight out of 10 cases they showed the ability of their approach to obtain close-to-native ligand binding poses within the 100 top-ranked complex models, with BGT being one of the unsuccessful cases. However, although near-holo conformations of receptor's structure were obtained with their workflow also for this target (featuring a $\text{RMSD}_{BS} < 2$ Å), no near-native ligand poses were retrieved within the top 100 models for this target, making it a well-suited test case to test our method. The second target is the recombinant ricin (hereafter RIC) [262], representative of proteins undergoing very minor but subtle conformational changes upon binding of its ligands (in this case neopterin, NEO) [263] (Figures 3.1-b,e). This is also testified by the RMSD_{BS} and ΔRoG_{BS} calculated over all the heavy atoms between the apo and holo experimental structures, respectively accounting to 1.0 Å and of -1% (Table 3.1). RIC belongs to the Astex Diverse Data Set [264] recently used to validate the AutoDockFR (AutoDock for Flexible Receptors) [265] docking software, in which receptor's flexibility is accounted by identifying a (sub)set of its side-chains to treat as flexible. Although AutoDockFR outperformed AutoDock Vina [266] in cross-docking experiments using receptor's apo conformations for most targets analysed in the

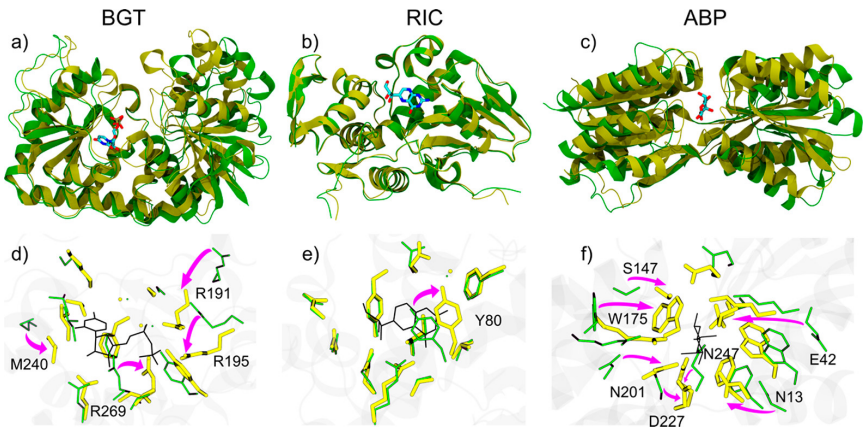

Figure 3.1: Comparison of the structural changes undergone by (a, d) BGT, (b, e) RIC, and (c, f) ABP upon binding of their ligands UDP, NEO, and ALL, respectively. (a-c) Overall conformational rearrangements of the proteins. The apo and holo proteins are shown in green and yellow ribbons, respectively, with the ligands in sticks coloured by atom type. (d-f) Detailed views of the local rearrangements occurring at the binding site. The conformations of residues lining the binding site in the apo and holo forms of the proteins are shown with thin green and thick yellow sticks, respectively, while the ligands are shown with thin black sticks and the protein is shown in transparent grey ribbons. The most significant reorientations upon ligand binding are indicated by magenta arrows. Image taken from ref. [22].

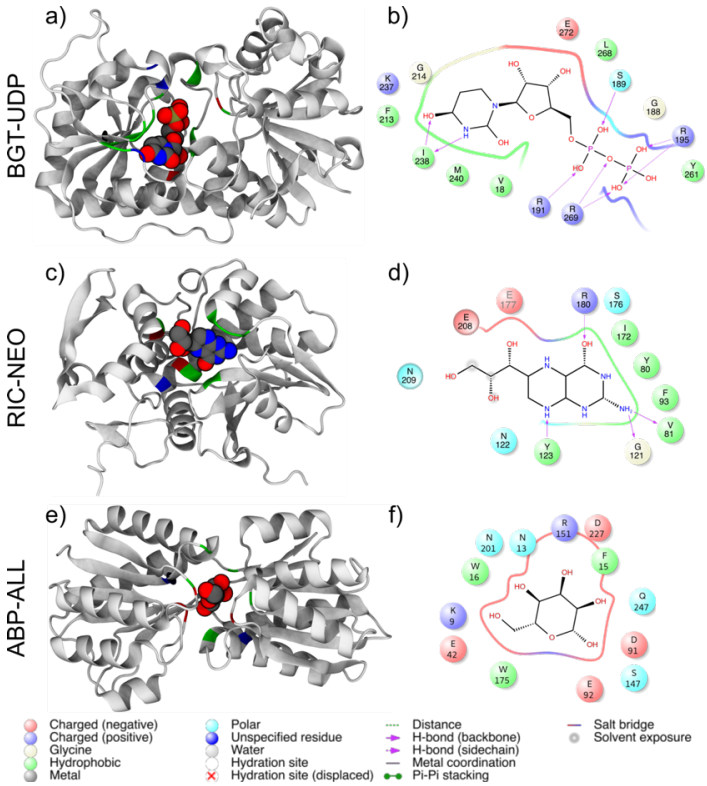

FIGURE 3.2: Location of the binding sites and additional structural details of the systems investigated in this work. a, c, e) Structures of the holo-proteins are shown as grey ribbons with residues lining the BS coloured by residue type and the ligands as spheres coloured by element; b, d, f) 2D-interaction diagrams for UDP, NEO and ALL with residues lining their respective BS, using the Ligand Interaction Diagram tool of Maestro [258]. Polar, apolar, negatively and positively charged residues are coloured cyan, green, red and blue respectively, with glycine residues coloured white. Image taken from ref. [22].

work, for RIC none of the aforementioned software was able to find any solution featuring a $\text{RMSD}_{lig} < 2.5$ Å from the experimental structure of the complex.

Finally, as third target we selected the allose binding protein (ABP) (Figures 3.1-c,f), also representative (like BGT) of targets undergoing extended hinge-bending motions upon binding of their ligands, in this case d-allose (ALL) [267]. In this case, RMSD_{BS} and ΔRoG_{BS} between the apo and holo experimental structures account respectively to 4.2 Å and of -26% (Table 3.1), making this target the one displaying the largest conformational changes among the three (for BGT, RMSD_{BS} and ΔRoG_{BS} are respectively 2.8 Å and of -10%). This protein therefore represents another good test case for our approach. Motta and Bonati [24] also selected this target to test their workflow based on ensemble docking calculations performed with Glide [268, 269] of its experimental ligand against target conformations generated through accelerated MD simulations [62]. Further details of the three systems investigated here concerning the RMSDs, the change in RoG between the apo and holo experimental structures and the characterisation of their binding sites are reported in table 3.1 and figure 3.2.

In the following we demonstrate that for all of the targets considered here, EDES is

able to generate druggable and holo-like receptor conformations. Moreover, using the widespread and freely available docking programs HADDOCK [270] and AutoDock4 [271], differing in their search and scoring schemes, on a set of EDES-generated conformations, we identify native-like ligand poses among the top-ranked ones for all the three systems studied. While being a proof of concept, this work opens the way to the automatic generation of holo-like conformations for a broad range of protein targets, representing a game-changing resource for structure-based drug design.

3.1.1 Methodological details

3.1.1.1 Standard MD Simulations

Standard all-atom MD simulations were carried out using the *pmemd* module of the AMBER16 [136, 272] molecular modeling software. Topology files were created for each system using the *LEaP* module of AmberTools17 starting from the experimental structures available in the Protein Data Bank (PDB IDs 1JEJ [259] and 1JG6 [259] for BGT and BGT-UDP, 1RTC [262] and 1BR5 [263] for RIC and RIC-NEO, and 1GUD [267] and 1RPJ [267] for ABP and ABP-ALL, respectively). The ff14SB [273] and GAFF [274] force fields were used for proteins and ligands, respectively. Missing parameters for the latter were generated using the antechamber module of AmberTools17. In particular, atomic restrained electrostatic potential charges were derived after a structural optimization performed with *Gaussian 09* [275]. Each structure was solvated with the explicit TIP3P water model [276], and its net charge was neutralized with the required number of randomly placed K^+ or Cl^- ions. The total number of atoms was $\approx$ 86.000 for BGT/BGT-UDP, $\approx$ 54.000 for RIC/RIC-NEO and $\approx$ 62.000 for ABP/ABP-ALL. Periodic boundary conditions were employed, and long-range electrostatics was evaluated through the particle-mesh Ewald algorithm using a real-space cutoff of 12 Å and a grid spacing of 1 Å per grid point in each dimension. The van der Waals interactions were treated by a Lennard-Jones potential using a smooth cutoff (switching radius 10 Å cutoff radius 12 Å). The initial distance between the protein and the edge of the box was set to be at least 16 Å in each direction. Multistep energy minimization with a combination of the steepest-descent and conjugate-gradient methods was carried out to relax internal constraints of the systems by gradually releasing positional restraints. Following this, the systems were heated from 0 to 310 K in 10 ns of constant-pressure heating (NPT) using the Langevin thermostat (collision frequency of 1 ps^{-1}) and the Berendsen barostat. After equilibration, four production runs of 2.5 μs each (for a total of 10 μs for each system) were performed for the apo systems, while a single 1 μs-long simulation was performed for each complex. A time step of 2 fs was used for pre-production runs, while equilibrium MD simulations were carried out with a time step of 4 fs in the NPT ensemble (using a MC barostat) after hydrogen mass repartitioning [144]. Coordinates from production trajectories were saved every 100 and 10 ps for MD_{apo} and MD_{holo}, respectively.

3.1.1.2 Metadynamics Simulations

Bias-exchange well-tempered metadynamics simulations [80, 152, 154, 171, 172] were performed on the three apo proteins using the GROMACS 2016.5 package

[247, 277] and the PLUMED 2.3.5 plugin [278]. The last conformation saved from the equilibration step from MD_{apo} was used as the starting structure for each simulation. AMBER parameters were ported to GROMACS using the *acpype parser* [279]. To enhance the sampling of different binding site shapes, we used the following four CVs defined by including all heavy atoms of the residues lining the binding site itself: the radius of gyration of the binding site (RoG_{BS}) calculated using the *gyration* built-in function of PLUMED and the numbers of (pseudo)contacts across the "inertia planes" ($CIP_{1,2,3}$) of the binding site, defined as the planes orthogonal to the three principal inertia axes and passing through the center of mass of the binding site. Binding site residues were defined as those within 3 Å (BGT and RIC) or 4 Å (ABP) of the ligand in the experimental structure of the complex (Table 3.2 and Figure 3.2). The cutoff was increased for ABP-ALL because of the low number of residues (seven) found when a 3 Å cutoff was used. Very similar definitions were found using the COACH-D Web server with the apo structures (Table 3.4). The CVs were calculated by an in-house tcl script based on the VMD orient function [1].

Protein	**Ligand**	$\mathbf{RMSD}_{prot}^{apo/holo}$ [Å]	$\mathbf{RMSD}_{BS}^{apo/holo}$ [Å]	$\mathbf{\Delta RoG}_{BS}^{apo/holo}$ [%]
BGT	UDP	2.3	2.8	-10
RIC	NEO	1.1	1.0	-1
ABP	ALL	4.7	4.2	-26

Table 3.1: RMSDs and RoGs for the system investigated in this work. For each system, ligand's name, RMSD of the protein/binding site and ΔRoG_{BS} between apo and holo experimental structures are reported. All values are calculated over all heavy atoms of the corresponding selection. Note that for all the three systems considered in this work, the RoG decreases upon ligand binding.

Namely, for each plane, residues lining the binding site were split into two lists A and B according to the positions of the geometrical centers of their backbones on each of the two sides of the inertia plane, and the overall number of (pseudo)contacts N_c between the two groups was calculated through the *coordination* keyword of PLUMED (see chap. 2.8 for details on EDES workflow).

Each of the four replicas of the bias-exchange metadynamics simulation was simulated for 100 ns (as our aim is primarily to enhance sampling of different shapes of the binding site and not to obtain converged free energy profiles), so that each window accumulated 400 ns of simulation time. Coordinates were saved every 10 ps. The height of the hills w was set to 0.6 kcal/mol for all systems, while it is customary to set the value of hills' width $\omega_{s,i}$ to be between 1/4 and 1/2 of the average fluctuations (standard deviations) of the relative CV during a short (of the order of a few hundreds of ps) plain MD run [170] (see chap. 2.4.2 for details on metadynamics). In our case, we set the widths to 0.15, 0.05, and 0.08 Å (RoG_{BS}), 5.4, 4.8, and 1.6 (CIP_1), 5.1, 3.2, and 4.9 (CIP_2), and 5.3, 3.1, and 6.0 (CIP_3) for BGT, RIC, and ABP, respectively. Hills were added every 2 ps, while the bias-exchange frequency was set to 20 ps. The bias factor for well-tempered metadynamics was

[1] A step by step tutorial on how to implement these CVs within the EDES framework can be found at the following link: `http://www.bonvinlab.org/education/biomolecular-simulations-2019/Metadynamics_tutorial/` referring to an hands-on tutorial that we gave during the BioExcel Summer School 2019. The files needed can be found here: `https://github.com/haddocking/EDES`.

set to 10. The "windows approach" briefly described in the the EDES workflow (see chap. 2.8) was implemented using RoG_{BS} as the driving parameter. We applied restraints (force constants set to 50 and 10 kcal/mol for the upper and lower walls, respectively, as we seek for compression rather than enlargement of the binding site) at values of RoG_{BS} that were 7.5% higher and lower than the value measured in the apo X-ray structure (RoG^{apo}_{X-ray}). Then, from the trajectory corresponding to this first window, we selected a random conformation of the protein whose RoG_{BS} was 5% lower than $RoG^{apo}_{X-ray}$and performed another simulation with walls centered at $\pm$ 7.5% RoG^{apo}_{X-ray} from this new center, repeating this procedure so as to simulate a total of four windows (Figure 3.3 and Table 3.3). It should be noted that the walls were set to allow partial overlap between adjacent windows, which indeed occurred in all cases (Figure 3.3).

Protein	**Binding site's residues**
BGT	V18,G188,S189,R191,R195,F213,G214, K237,I238,M240,Y261,L268,R269,E272
RIC	Y80,V81,F93,G121,N122,Y123,I172,S176, E177,R180,E208,N209
ABP	K9,N13,F15,W16,E42,D91,E92,S147, R151,W175,N201,D227,Q247

Table 3.2: The residues lining the BS for the system investigated in this work, defined exploiting the experimental holo structure.

	RoG_{BS} [Å]			
System	$\mathbf{C}_{w1}$	$\mathbf{C}_{w2}$	$\mathbf{C}_{w3}$	$\mathbf{C}_{w4}$
BGT	9.87	9.38	8.88	8.39
RIC	7.40	7.04	6.67	6.31
ABP	9.70	9.21	8.73	8.24

Table 3.3: Details of EDES implementation for the systems investigated in this work. C_{wi} indicates the value of the RoG_{BS} (Å) at which the ith window was centered.

System	BS definition	BS residues	BS fraction
BGT	X-ray	18,188,189,191,195,213,214,237,238,240,261,268,269,272	
	COACH-D	18,188,189,195,213,214,237,238,240,243,261,268,269,272 (2)	13/14
RIC	X-ray	80,81,93,121,122,123,172,176,177,180,208,209	
	COACH-D	79,80,81,93,121,122,123,172,176,177,180 (1)	10/12
ABP	X-ray	9,13,15,16,42,91,92,147,151,175,201,227,247	
	COACH-D	9,13,15,16,91,92,145,147,151,175,201,227,247 (1)	11/13

Table 3.4: Comparison between the compositions of the BS as defined in the main text (BS_{X-ray}) and those identified through the COACH-D webserver [257]. The rank of the site identified by COACH-D is reported in parenthesis in the third column after the list of residues. The last column shows the number of residues identified by COACH-D with respect to the selection based on the experimental holo structure. It shows that COACH-D recovers more than 80% of the residues lining BS_{X-ray} for all three cases.

3.1.1.3 Cluster analysis

The multi-step clustering analysis was performed using an in-house set of bash scripts interfacing with the R software [256], according to the following scheme (see chap. 2.7.2 for theoretical details on the cluster analysis) and employing the four CVs used for the sampling as clustering parameters:

1. The RoG distribution of MD trajectories is binned into a pre-defined number of 10 equally-wide slices. This is achieved by giving in input to the script the desired width (in Å) of each slice;

2. The *hclust* module of the *cluster* R-package was used to perform a hierarchical agglomerative clustering (HAC) [210, 212–215] with the Ward method [249, 250] in which the clusters are extracted separately from each slice. The representative of each cluster was selected as the one closest to the center, according to the euclidean metrics calculated on the CVs. The number of generated clusters within each slice was set to $x_i = (N_i/N_{tot}) \cdot N_c$, where N_i, N_{tot}, and N_c are, respectively, the number of structures within the ith slice, the total number of structures in the simulation, and the total number of clusters. For this work, N_c was set to 500.

3. The 500 cluster representatives extracted in the previous step are used as starting points for a subsequent K-means [210, 213, 216–219] cluster analysis performed on the whole trajectory and generating the same number of clusters. This analysis was performed with the *k-means* module of the *cluster* R-package, setting to 10.000 the maximum number of iteractions and using the euclidean metrics.

4. The 500 final clusters extracted from the K-means step are selected for ensemble-docking calculations.

However, together with using this clustering protocol in the framework of ensemble-docking calculations, we also assessed its performance with respect to (i) the usage of random-initialized K-means clustering without the previous HAC step; (ii) the usage of a different number of selected clusters for docking and (iii) the metrics

used. In particular, for the latter case we compared the results obtained with our protocol with the ones obtained with a standard approach based on the RMSD metrics. Results of this assessment will be discussed in chap. 3.1.3.

3.1.1.4 Molecular Docking calculations

Molecular docking calculations were performed with AutoDock4 [271] and the HADDOCK Web server version 2.2 [270, 280]. This choice allowed our methodology to be validated against two programs differing in their search algorithms, scoring functions, and pose selection schemes. However, before being used with EDES-generated conformations, the performance of both software was first validated in re-docking experiments against experimental holo structures (Table 3.5).

System		AutoDock4		HADDOCK		
Protein	**Simulation**	**Rank**	**RMSD**$_{lig}$	**Rank**	**RMSD**$_{lig}$	**F**$_{nat}$
BGT	Xray$_{apo}$	- (-)	8.8 (9.6)	- (-)	3.16±0.33 (3.12±0.57)	0.54±0.03 (0.60±0.09)
	Xray$_{holo}$	1 (1)	0.3 (0.9)	1 (1)	0.39±0.12 (1.56±0.21)	0.97±0.02 (0.91±0.04)
RIC	Xray$_{apo}$	- (-)	7.3 (3.7)	- (-)	3.48±0.05 (3.88±0.26)	0.53±0.00 (0.53±0.05)
	Xray$_{holo}$	1 (1)	0.7 (1.8)	1 (1)	0.59±0.08 (1.19±0.06)	0.93±0.00 (0.93±0.00)
ABP	Xray$_{apo}$	- (-)	15.0 (15.7)	8 (49)^a	1.9±0.2 (1.59)^a	0.47±0.00 (0.47)^a
	Xray$_{holo}$	1 (1)	0.5 (0.8)	1(1)	0.68±0.13 (0.73±0.06)	1.0±0.0 (1.0±0.0)

Table 3.5: Performance of AutoDock4 and HADDOCK in docking the experimental structures of UDP, NEO and ALL onto the X-ray structures of BGT, RIC and ABP respectively. Values refer to RMSD$_{lig}$ (in Å) from to the experimental pose in the X-ray holo-structures as obtained in rigid docking calculations after alignment of the BS. The fraction of native contacts F_{nat} is defined as the number of native intermolecular contacts identified in a docking pose divided by the total number of contacts in the reference structure. The contacts are evaluated within a shell of 5 Å from the ligand in the experimental structures. Values in parentheses were obtained using a flexible ligand model. Ranking is reported only when RMSD$_{lig}$ is lower than 2 Å. The HADDOCK statistics and rankings are based on the average over the top 4 poses of the first acceptable (RMSD$_{lig} \leq 2$ Å) cluster.
^{a}: Cases in which no acceptable clusters were obtained. In this case the reported statistics corresponds to single pose statistics for the first acceptable (≤ 2 Å) and best (between brackets) poses.

For this work, calculations with both programs were performed using their default settings, apart from the following changes: In AutoDock4, we used the Lamarckian genetic algorithm (LGA) to perform a hybrid global-local search of the docking poses. The grid density (spacing parameter changed from 0.375 to 0.25 Å) and number of energy evaluations (ga-num-evals increased by a factor of 10 from the default value) were both increased, with the purpose to avoid repeating each calculation several times to obtain converged results. For each target, 500 independent rigid docking calculations (one with each conformation from the ensemble) were performed using an adaptive grid enclosing all of the residues belonging to the binding site. Next, the top poses (in total 500, one for each docking run) were clustered using the *cpptraj* [246] module of AmberTools17 with a hierarchical agglomerative algorithm (with the "complete" method) and a cutoff of 1.5 Å for the $RMSD_{BS}$ distance matrix. In HADDOCK, a single docking run was performed per case, starting from the various ensembles of 500 conformations, with increased sampling (10000/400/400 models for

rigid-body docking (*it0 step*), semi-flexible refinement (*it1 step*), and final refinement in explicit solvent (*wat step*)[2]. The weight of the intermolecular van der Waals energy for the initial rigid-body docking stage was increased to 1.0 (from the default value of 0.01), RMSD-based clustering was selected with a cutoff of 1 Å, and the docking was guided by ambiguous distance restraints defined for the residues of the binding site and the ligand [202]. In the rigid-body stage, binding site residues were defined as active, effectively drawing the ligand into the binding site without restraining its orientation. For the subsequent stages, the restraints were such that only the ligand was active, allowing it to explore the binding site better while maintaining at least one contact with its interacting residues.

3.1.2 Sampling of holo-like conformations

In this section we discuss the performance of EDES in generating holo-like geometries of the three proteins. We start with BGT, in which the binding event induces large conformational rearrangements in the BS, particularly in the orientation of three (positively charged) arginine residues (R191, R195, and R269) neutralising the negative charge of the diphosphate group of the ligand UDP (Figures 3.1-a,d and 3.2). Figure 3.4 shows the performance in sampling holo-like conformations of the different simulation approaches presented here: the unbiased simulations of the apo and holo systems, respectively MD_{apo} and MD_{holo}, and EDES approaches (see chap. 2.8 for EDES workflow and chap. 3.1.1 for methodological details). The performance is evaluated by means of the RMSD metrics, calculated over all the heavy atoms for the binding-competent region ($RMSD_{BS}$), the one affecting the most docking results.

[2]General details on HADDOCK docking protocol can be found in ref. [270].

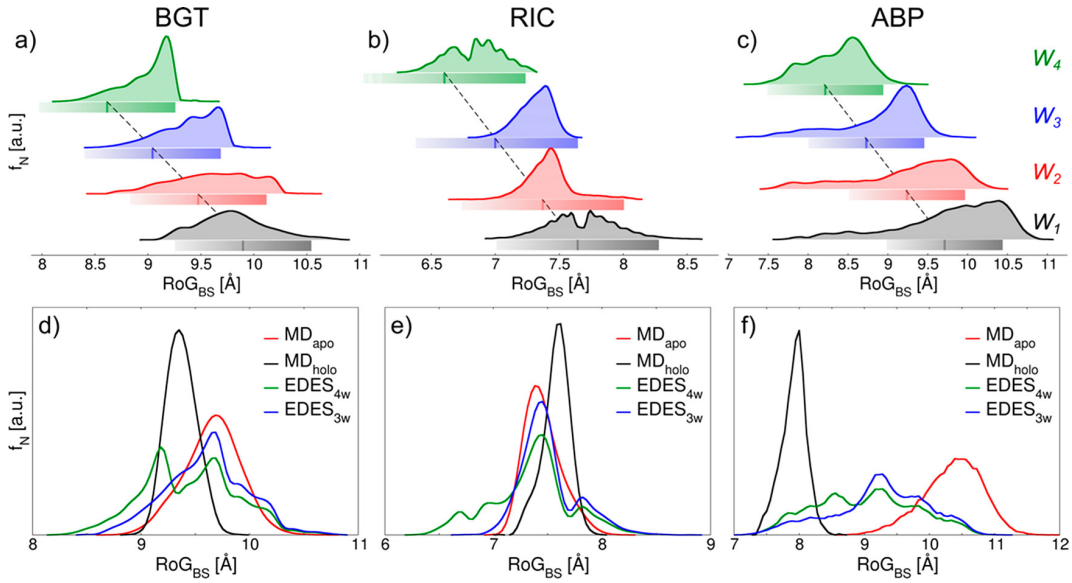

Figure 3.3: Distributions of RoG_{BS} values. (a-c) Distributions from each EDES window for (a) BGT, (b) RIC, and (c) ABP. The colored bar below each distribution indicates the position of the lower and upper walls set for RoG_{BS} in that window, and the color gradient indicates a higher force constant for the upper wall than the lower wall (the centers of the windows are indicated by darker lines within the bars and are connected by a black dashed line). (d-f) Comparison of RoG_{BS} normalised distributions (area under each curve equal to 1, bin size set to 0.1 Å) obtained from the different simulations performed in this work. Image taken from [22].

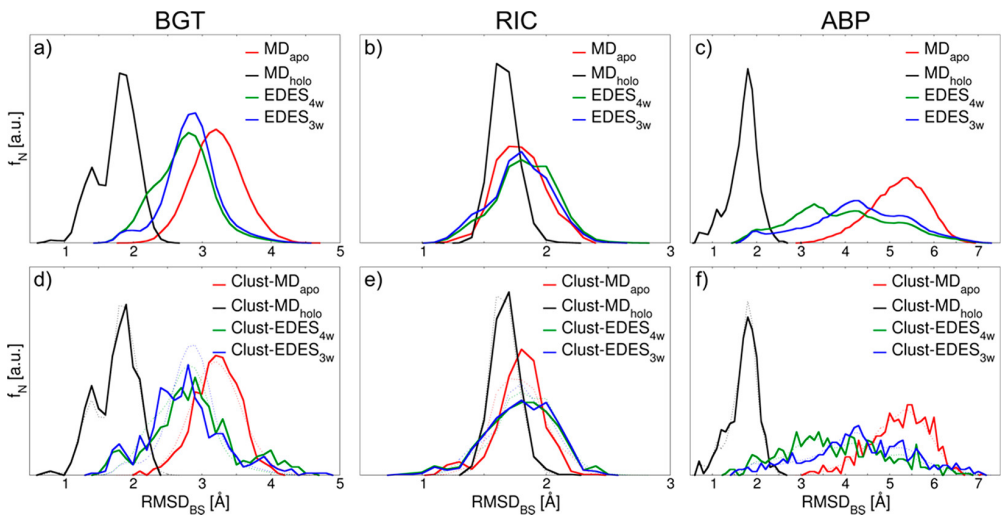

Figure 3.4: Normalised distributions (area under each curve equal to 1, bin size set to 0.1 Å) of the RMSD of the binding site heavy atoms ($RMSD_{BS}$) with respect to the holo structure. The upper (a-c) and lower (d-f) rows show the distributions calculated over all of the snapshots extracted from each MD simulation and over cluster representatives only, respectively. The faint dotted lines in (d-f) correspond to the distributions in (a-c). Image taken from ref. [22].

Inspecting figure 3.4 a very poor overlap is revealed between MD_{apo} and MD_{holo} distributions. This was partially expected, as a relatively short plain MD simulation is generally unable to sample extended conformational changes. MD_{holo} clearly performs extremely well but it is used in this work purely as reference distribution, as the knowledge the structure of the complex will be typically unavailable in real applications. On the other hand, EDES distributions are centred somewhat in between the ones obtained from the unbiased MD simulations. In addition, most conformations sampled by EDES have $RMSD_{BS}$ lower than 2.8 Å from the experimental structure of the complex. This RMSD value is also the one between the experimental apo and holo structures of the target (see Table 3.1), showing that EDES is able to reproduce closer-to-holo states with respect to the experimental apo geometry. Moreover, EDES distributions reveal a shoulder at lower RMSDs that increases the percentage of conformations with $RMSD_{BS} \leq 2$ Å compared with MD_{apo}, a feature that persists after clustering (Figure 3.4). Results thus show that our protocol is able to generate conformations with a (partially) collapsed binding site also in absence of ligand interactions (triggering such collapse), as also testified by the inspection of figures 3.3 and 3.5, displaying respectively RoG and CIPs distributions. The effect of the enhanced sampling with respect to the unbiased MD_{apo} is particularly evident in the case of arginine R269, pointing to the center of the binding pocket in the experimental apo conformation: while in MD_{apo} its side-chain remains in the center of the binding site, a large fraction of EDES-generated structures features a displaced R269, making room for ligand binding (Figure 3.6). Furthermore, the ability of our approach to produce near-holo conformations is also proved within the CIP metrics, describing the 3D shape of the pocket. Indeed, figure 3.7 makes evident the improved overlap between MD_{holo} and EDES distributions with respect to MD_{apo}. In particular, only EDES samples conformations with CIP values virtually identical to those of the experimental holo structure (black sphere in Figure 3.7).

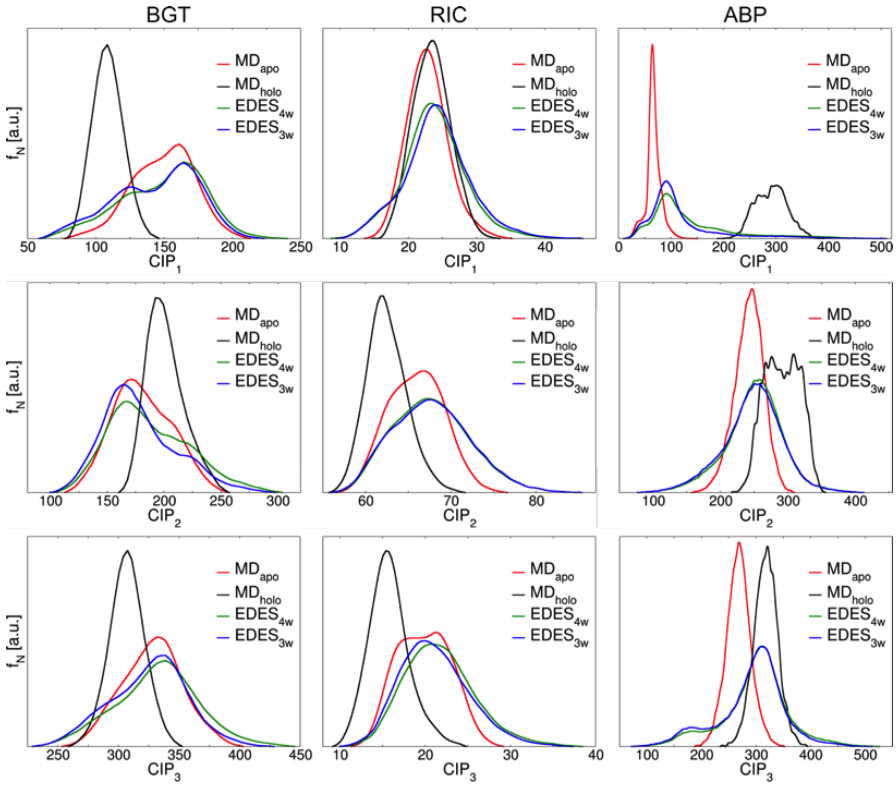

Figure 3.5: Normalised distributions of the "(pseudo)contacts across inertia planes" (CIPs) collective variables values sampled during the various MD simulations performed in this work for BGT, RIC and ABP. Image taken from ref. [22].

Finally, concerning the clustering protocol, acting specifically on the 3D shape of the pocket (and not on coarse-grained descriptors such as the RMSD) it produced cluster distributions with an increased percentage of native-like BS geometries, compared with the fraction sampled during MD simulations (Table 3.6).

However, as already pointed out, our methodology relies on the accurate identification of the binding site(s). For this reason, we also investigated its performance using a slightly different definition of the binding site. Namely, we took advantage of one of the many available site-finding web servers, COACH-D [257] to determine consensus binding sites of all three targets (Table 3.4), resulting in a very high overlap between the experimentally derived sites and those identified by COACH-D. For example, in the case of BGT, the top identified BS shares 13 out of 16 residues with the experimental holo structure. We thus repeated all of the EDES simulations for the three systems using the alternative definition of the binding site obtained from COACH-D, obtaining virtually identical results (Table 3.7 shows the results for BGT) to the ones obtained with the experimental BS definition. This indicates that our method is not sensitive to the exact binding site definition and so it can also be used successfully in conjunction with site detection algorithms in cases where only an apo structure is known. This aspect will be more deeply addressed in the second work presented here (chap. 3.2).

Our method has been primarily developed for targets undergoing rather large

conformational changes, so for the ones for which unbiased MD simulations are generally not sufficient to generate near-holo geometries. However, in several real cases, the extent of apo-to-holo conformational rearrangement(s) could be unknown. For this reason, we decided to test our method also with RIC, a paradigmatic protein for cases in which the binding event is associated only to minor conformational changes (Figure 3.2-b,e), in the spirit of validating the protocol also for such cases. Furthermore, previous works (see e.g. ref. [24]) have already pointed out that not always enhanced-sampling approaches improve the percentage of holo-like conformations generated with respect to standard MD simulations, in particular when the target only undergoes small conformational rearrangements upon binding. However, although small, RIC rearrangements were hardly handled by algorithms exploiting the flexibility of the binding site through side-chain torsional angles [265]. For this reason, RIC has proven to be a very difficult target for both rigid and flexible docking calculations starting from the apo X-ray structures using both AutoDock Vina [266] and the recently introduced AutoDockFR [265] software (see Table 1 in ref. [265]).

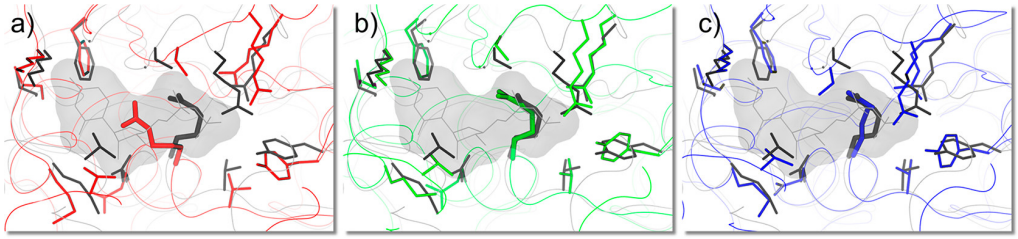

FIGURE 3.6: Binding site views of the lowest-RMSD conformations of BGT with respect to the bound complex extracted from (a) MD_{apo}, (b) $EDES_{4w}$, and (c) $EDES_{3w}$. UDP is shown in gray lines and transparent surface. The proteins are shown as gray (holo experimental structure), red (MD_{apo}), dark green ($EDES_{4w}$, and blue $EDES_{3w}$) thin ribbons, with side chains of residues lining the binding site represented as sticks, which are thicker for R269. Image taken from ref. [22].

In contrast, in our case, standard and enhanced MD simulations were able to reproduce near-holo geometries of its binding pocket (although the enhanced approach retrieved conformations closer to the holo structure than those obtained from $MD_{apo/holo}$; see Table 3.6 and Figure 3.4). The good performance of the unbiased approach of the apo system in this case was expected, due to the minor extent of conformational changes involved in the binding, while EDES confirmed its ability to sample near-holo states also for such targets. This trend is evident also in reproducing holo-like conformations of the entire protein (Figure 3.8), as both (un)biased approaches sampled a relatively large fraction of such structures. On the basis of these results, we are confident that our approach could also be applied for other cases in which targets undergo minor conformational changes upon ligand binding. As stated above, this is particularly encouraging since in a real case one might not know the extent of the conformational change in advance. The third protein considered in this work is ABP (Figure 3.2-c,f), undergoing the largest conformational changes among the other cases considered here, upon binding of its (small and neutral) ligand ALL. For this system, the performance difference between the unbiased and biased approaches is the most evident (Figures 3.3 and 3.4), as

the unbiased approach is unable to produce any single conformation featuring an RMSD_{BS} lower than 2 Å from the holo X-ray structure (Table 3.6). Moreover, also in this case, despite only directly enhancing the sampling of the binding region, our approach is also able to drag the whole protein structure toward close-to-holo conformations (Figure 3.8).

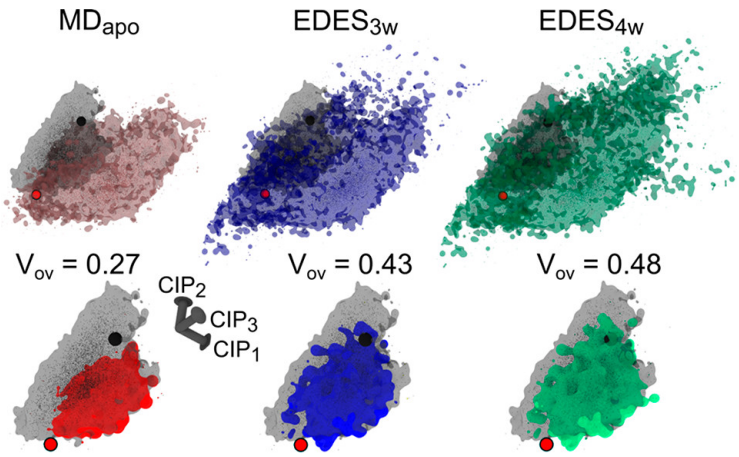

Figure 3.7: Sampling of the 3D space defined by CIP_1, CIP_2, and CIP_3 during the MD simulations of BGT and BGT–UDP. *Top row*: Comparison of the MD_{apo} (red), $EDES_{3w}$ (blue), and $EDES_{4w}$ (green) distributions with the MD_{holo} distribution (dark gray). The distributions are shown both as solid points and as transparent surfaces. The locations of the apo and holo structures are indicated by red and black spheres, respectively. *Bottom row*: Envelopes of the MD_{apo} (red), $EDES_{3w}$ (blue), and $EDES_{4w}$ (green) distributions overlapping with the MD_{holo} distribution (shown in dark gray as a reference). Also reported are the volumes of the overlapping distribution, Vov (estimated with Voss Volume Voxelator (http://3vee.molmovdb.org) using a probe radius of 3 Å). Image taken from ref. [22].

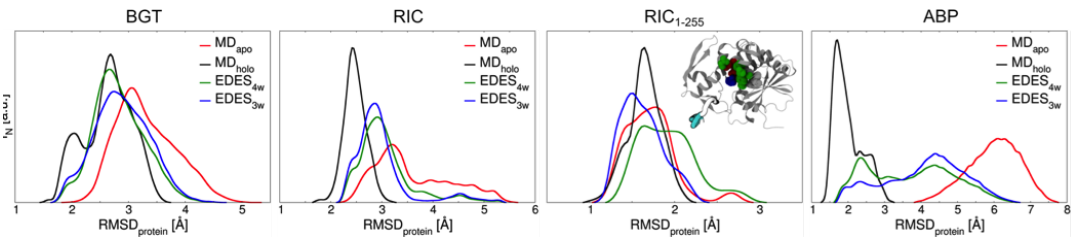

Figure 3.8: Normalised distributions (area under each curve normalised to 1; bin size equal to 0.1 Å) of the RMSD of the protein backbone ($RMSD_{protein}$) with respect to the holo structure for BGT, RIC, RIC excluding the flexible C-terminal loop (residues 256 to 267, displayed as tick ribbon in the inset showing the protein as gray ribbons and the residues of the BS as spheres colored by residue type), and ABP. Image taken from [22].

System		$RMSD_{BS} < 1.5$ [Å]		$RMSD_{BS} < 2.0$ [Å]	
Protein	**Simulation**	**Trajectory**	**Clusters**	**Trajectory**	**Clusters**
BGT	MD_{apo}	-	-	0.06 (1.69)	-
	MD_{holo}	23.9 (0.75)	26.6 (0.75)	85.3 (0.75)	84.8 (0.75)
	$EDES_{3w}$	0.02 (1.31)	0.4 (1.31)	3.8 (1.31)	4.0 (1.31)
	$EDES_{4w}$	0.02 (1.31)	0.2 (1.31)	5.0 (1.31)	8.4 (1.31)
RIC	MD_{apo}	9.7 (1.00)	9.6 (1.10)	89.4 (1.00)	93.2 (1.10)
	MD_{holo}	13.0 (0.91)	12.4 (0.95)	99.6 (0.91)	97.8 (0.95)
	$EDES_{3w}$	16.5 (0.77)	17.6 (0.81)	85.5 (0.77)	85.4 (0.81)
	$EDES_{4w}$	13.0 (0.77)	19.0 (0.81)	81.4 (0.77)	82.4 (0.81)
ABP	MD_{apo}	-	-	-	-
	MD_{holo}	25.8 (0.48)	25.6 (0.67)	96.4 (0.48)	88.6 (0.67)
	$EDES_{3w}$	0.5 (1.17)	2.4 (1.20)	6.1 (1.17)	8.4 (1.20)
	$EDES_{4w}$	0.6 (1.17)	1.4 (1.20)	8.6 (1.17)	9.4 (1.20)

Table 3.6: Performance of different MD simulation protocols in reproducing native-like conformations of the BS of BGT, RIC, and ABP. The performance is measured by the percentage of conformations with $RSMD_{BS} \leq 1.5$ or 2 Å with respect to the experimental structure. The headings *trajectory* and *clusters* refer to snapshots extracted from the full trajectories and to cluster representatives, respectively. The lowest value of $RMSD_{BS}$ (in Å) is reported in parentheses. As expected, this percentage is high for MD_{holo}. While a very low number of such conformations was sampled in MD_{apo}, a large fraction was obtained by EDES using either three or four windows.

System		$\text{RMSD}_{BS} < 1.5$ [Å]		$\text{RMSD}_{BS} < 2.0$ [Å]	
Protein	**Simulation**	$\mathbf{BS}_{Xray}$	$\mathbf{BS}_{COACH}$	$\mathbf{BS}_{Xray}$	$\mathbf{BS}_{COACH}$
BGT	$EDES_{4w}$	0.02 (1.31)	0.01 (1.41)	5.0 (1.31)	6.4 (1.41)
	$EDES_{3w}$	0.02 (1.31)	0.01 (1.41)	3.8 (1.31)	4.1 (1.41)

Table 3.7: Performance of EDES in reproducing native-like conformations of the BS of BGT using either the BS_{Xray} or the BS_{COACH} definitions. The performance is measured by the percentage of conformations featuring a value of RMSD_{BS} lower than 1.5 or 2 Å. The lowest value of RMSD_{BS} (Å) with respect to the holo X-ray structures is reported in parentheses.

3.1.3 Impact of the cluster analysis

As already pointed out, the outcome of MD-based ensemble-docking is directly linked to the ability to include, in the pool of receptor conformations, also geometries prone to host the ligand(s) to be docked. For this reason, as in general the true bound-like receptor conformation for the ligand of interest is not known a priori, a cluster analysis on MD trajectories is performed, in order to extract a certain number of conformation representatives. Clearly, in absence of information concerning the complex (such as the RoG of its putative binding site(s)), aim of the cluster analysis is to maximise the structural diversity of the ensemble of clusters, at least at the binding site.

In our case, despite not making any use of specific knowledge of the structure of the complexes, our informed strategy was able to generate a larger fraction of cluster structures displaying $\text{RMSD}_{BS} < 2$ Å than that obtained from the standard application of K-means using randomly selected conformations as starting points (Figure 3.9). This is not surprising, as the random initialization strategy is considered one of the most unreliable ones [255] (see chap. 2.7.2 for details).

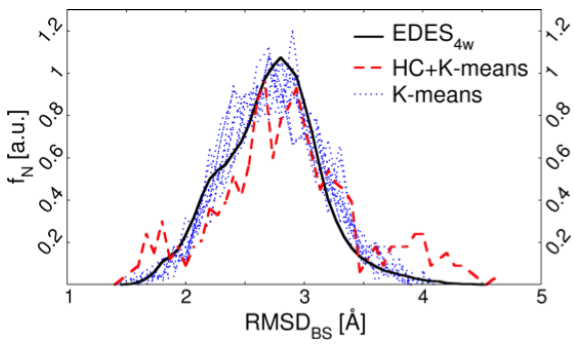

Figure 3.9: Performance of the multi-step clustering strategy adopted in this work (hierarchical clustering -HC- within slices of the RoG_{BS} distribution followed by a second clustering with K-means starting from the points obtained at the first step) in selecting cluster representatives featuring a low RMSD_{BS} from the ligand-bound structure. The black line represents the RMSD distribution obtained from the EDES_{4w} trajectory of the BGT protein. It appears how our approach (red dashed line) results in more uniform distributions with higher tail populations compared to standard cluster analysis performed with K-means only without dividing in slices the whole trajectory (10 independent cluster analyses were performed, whose distributions are shown in blue dots). Image taken from ref. [22].

Furthermore, we also tested our clustering protocol against the number of generated clusters, aiming to understand the smallest number of clusters still yielding to include a suitable fraction of near-holo conformations in the ensemble of clusters. To do so, we performed additional cluster analyses, with the protocol we developed, in which we decreased the total number of clusters from 500 to 200 and 100. Using the resulting sets of conformations, we performed additional docking calculations with AutoDock4 [271]. The results, reported in table 3.8, reveal that the fraction of holo-like conformations (featuring $\text{RMSD}_{BS} \leq 2$ Å) remains virtually constant when going from 500 to 100 clusters and that also docking performance is comparable when different numbers of starting conformations are used (Table 3.9).

Finally, another aspect that we decided to investigate is the impact of the metrics used for the clustering. We performed this analysis on the BGT–UDP system, for which we compared the performance in selecting holo-like conformations from an MD trajectory of our clustering protocol (CA_{CVs}) to a more standard analysis based on the distance RMSD (dRMSD) of the binding site (CA_{dRMSD}) using the hierarchical agglomerative method implemented in the *cpptraj* [246] module of the AMBER [136] package. Moreover, we performed docking calculations with AutoDock4 [271] against the conformations extracted with both approaches to compare the impact of these clusters on docking performance. This confirmed that, as highlighted by others [24, 220], clustering strategies based on a 3D descriptor linked to the pocket shape appear to be more effective in delivering maximally different conformations of the binding site than those based on the RMSD of the putative region (Table 3.10). The impact of the clustering strategy used is even more evident when referring to docking results, with no near-native pose within the top 10 when using structures obtained from CA_{dRMSD} clustering (Table 3.11).

System		Number of clusters		
Protein	**Simulation**	**100**	**200**	**500**
BGT	$EDES_{4w}$	9.0 (1.65)	8.0 (1.65)	8.4 (1.31)
	$EDES_{3w}$	4.0 (1.62)	4.0 (1.49)	4.0 (1.31)
RIC	$EDES_{4w}$	78.0 (0.89)	76.5 (0.89)	77.8 (0.81)
	$EDES_{3w}$	80.0 (0.89)	80.0 (0.89)	77.0 (0.81)
ABP	$EDES_{4w}$	10.0 (1.27)	8.0 (1.20)	9.4 (1.20)
	$EDES_{3w}$	7.0 (1.20)	7.5 (1.20)	8.4 (1.20)

Table 3.8: Performance of the multi-step clustering protocol employed in this work in reproducing native-like conformations of the BS at varying number of clusters. Reported are the percentages of conformations featuring $RSMD_{BS}$ values lower than 2 Å as well as the lowest value of the RMSD (Å) with respect to the holo X-ray structures (in parentheses).

		Number of clusters					
System		**100**		**200**		**500**	
Protein	Simulation	$EDES_{3w}$	$EDES_{4w}$	$EDES_{3w}$	$EDES_{4w}$	$EDES_{3w}$	$EDES_{4w}$
BGT	Sampl. perf. [%]	2.0	3.0	1.0	2.0	1.8	2.0
	Pose rank	1 (0.7)	1 (1.3)	8 (0.9)	2 (1.5)	1 (0.6)	2 (1.5)
RIC	Sampl. perf. [%]	2.0	3.0	1.5	2.0	2.6	3.4
	Pose rank	2 (0.8)	1 (0.8)	3 (0.8)	1 (0.9)	4 (0.7)	1 (0.9)
ABP	Sampl. perf. [%]	4.0	1.0	3.0	1.5	3.0	3.2
	Pose rank	4 (1.2)	5 (1.8)	3 (1.6)	10 (1.4)	2 (0.8)	2 (0.7)

Table 3.9: Performance of AutoDock4 in reproducing the experimental structures of the BGT-UDP, RIC-NEO, and ABP-ALL complexes in ensemble-docking calculations with different number of clusters. Results refer to clusters of docking poses obtained from a cluster analysis performed on all generated complexes using as metrics the dRMSD with a cutoff of 1.5 Å. The sampling performance is calculated as the percentage of poses within 2 Åfrom the native structure out of all the top poses considered for each ensemble of receptor structures. The ranking of the first native-like pose obtained using the highest score within each cluster is also reported for each case, with the values of $RMSD_{lig}$ in parentheses.

System		$RMSD_{BS} < 1.5$ [Å]		$RMSD_{BS} < 2.0$ [Å]	
Protein	**Simulation**	CA_{CVs}	CA_{dRMSD}	CA_{CVs}	CA_{dRMSD}
BGT	$EDES_{4w}$	0.2 (1.31)	-	8.4 (1.31)	4.4(1.51)
	$EDES_{3w}$	0.4 (1.31)	-	4.0 (1.31)	2.8 (1.64)

Table 3.10: Performance of EDES in reproducing native-like conformations of the BS of BGT using 500 cluster structures extracted either with our multi-step protocol (CA_{CVs}) or through a single-dimensional cluster analysis based on the dRMSD of the BS (CA_{dRMSD}). The performance is measured by the percentage of conformations featuring a value of $RSMD_{BS}$ lower than 1.5 or 2 Å. The lowest value of the RMSD (Å) with respect to the holo X-ray structures is reported in parentheses.

	$EDES_{3w}$		**$EDES_{4w}$**	
	CA_{CVs}	CA_{dRMSD}	CA_{CVs}	CA_{dRMSD}
Sampl. perf. [%]	1.8	0.8	2.0	2.0
Pose rank	1 (1)	12 (18)	2 (2)	16 (21)
Clus. Pop.	4	1	4	1
$RMSD_{lig}$ [Å]	0.6 (0.9)	0.6 (0.7)	1.5 (1.2)	0.7 (0.7)

Table 3.11: Performance of AutoDock4 in reproducing the experimental structures of the BGT-UDP in ensemble-docking calculations using 500 structures extracted with different clustering strategies from $EDES_{3w}$ and $EDES_{4w}$ simulations. Results refer to clusters of docking poses obtained from a cluster analysis performed on all generated complexes using as metrics the dRMSD of the binding site with a cutoff of 1.5 Å. The sampling performance (third row) is calculated as the percentage of poses within 2 Å of the native structure out of all the top poses considered for each ensemble of receptor structures. The fourth row reports the ranking of the first native-like pose obtained using the highest score within each cluster for ranking. In parentheses, the rank of the same cluster is reported when the average score over the top three poses is used instead. The fifth row reports the population of the corresponding cluster in the same column. The last row reports the average heavy-atoms RMSD of the ligand calculated for the top cluster, with standard deviation in parentheses.

3.1.4 Docking performance

In this subsection we describe the results of docking calculations performed for each target with the EDES-generated ensemble of conformations. Specifically, in the case of AutoDock4, for each target, 500 independent runs were carried out, one for each different receptor conformation in the ensemble. Next, a cluster analysis on the top poses obtained from each individual docking run is performed, to obtain the final results presented here. On the other hand, in HADDOCK all of the 500 conformations within each ensemble were used in a single docking run. The results are reported in tables 3.12, 3.13 and 3.14, while details on the molecular docking protocol can be found in chap. 3.1.1.4.

For the BGT–UDP complex, both software achieved improved performance (defined as the percentage of ligand poses displaying a value of RMSD_{lig} lower than 2 Å from the holo structure) when coupled to EDES rather than MD_{apo} (Tables 3.12, 3.13 and 3.14), generating consistent fractions of native-like ligand poses. Specifically AutoDock4 and HADDOCK, respectively produced up to 2% and 14% of native-like poses when coupled to EDES and 0% (for AutoDock) 2% (for HADDOCK) when the clusters from MD_{apo} are used. Another important feature is that both programs were able to rank at least one native-like pose among the top two when coupled with EDES, regardless of the number of windows used to generate the clusters of conformations (see Tables 3.12, 3.13 and 3.14 and Figure 3.10). This highlights that also docking performance is not very sensitive to the exact number of windows used, reflecting the trend already observed for the generation of holo-like conformations. In the case of RIC-NEO, on the other hand, results clearly show that using clusters coming from both the unbiased and biased approaches result in the generation of native-like ligand poses within the top scored ones. As already mentioned, however, the good performance of the enhanced approach was not guaranteed and confirms the possibility to use our workflow also for targets undergoing minor conformational changes upon binding. Finally, also in the case of the ABP-ALL system, we obtained very encouraging results. In this case, we observe that HADDOCK's performance is overall better with EDES_{3w}. In fact, in the case of EDES_{4w}, the top pose obtained with HADDOCK, although satisfying the $\text{RMSD}_{lig} \leq 2$ Å criterion, has a flipped orientation, confirming the known limitation of using the RMSD criterion alone to evaluate the docking performance. However, still in the case of the EDES_{4w} ensemble, a ligand pose (almost) perfectly overlapping with the experimental one is retrieved in the 10th cluster ($\text{RMSD}_{lig} = 0.9 \pm 0.1$ Å). A final word should be spent on the fact that for both BGT and RIC the results were virtually independent of the number of EDES windows used, trend which was verified for both docking software used. Clearly, further studies will be needed to optimise the number and the width of the windows used, possibly exploiting a set of intrinsic properties of each protein so as to set up target-dependent rules.

		$\mathbf{MD}_{apo}$	$\mathbf{MD}_{holo}$	$\mathbf{EDES}_{3w}$	$\mathbf{EDES}_{4w}$
sampling performance [%]	BGT-UDP	-	84.6	1.8	2.0
	RIC-NEO	3.8	10.8	2.6	3.4
	ABP-ALL	-	94.0	3.0	3.2
pose rank	BGT-UDP	-	1 (1)	1 (1)	2 (2)
	RIC-NEO	1 (1)	1 (1)	4 (3)	1 (1)
	ABP-ALL	-	1 (1)	2 (2)	2 (7)
cluster population	BGT-UDP	-	48	4	4
	RIC-NEO	15	20	10	13
	ABP-ALL	-	225	9	5 (6)
RMSD_{lig} [Å]	BGT-UDP	-	1.2 (0.7)	0.6 (0.9)	1.5 (1.2)
	RIC-NEO	1.0 (0.9)	0.6 (0.6)	0.7 (0.7)	0.9 (0.8)
	ABP-ALL	-	0.7 (0.2)	0.8 (0.3)	0.7 (0.2)

Table 3.12: Performance of AutoDock4 in reproducing the experimental structures of the BGT-UDP, RIC-NEO, and ABP-ALL complexes in ensemble docking calculations. Results refer to clusters of docking poses obtained from a cluster analysis performed on all generated complexes (500 for each ensemble of clusters of receptor structures, corresponding to the top pose from each independent docking run for that ensemble) using the distance-RMSD (dRMSD) with a cutoff of 1.5 Å as metrics. The sampling performance is calculated as the percentage of poses within 2 Å from the native structure out of the top poses considered for each ensemble of receptor structures. The pose rank refers to the ranking of the first native-like pose obtained using the highest score within each cluster as sorting criterion. In parentheses, the rank of the same cluster is reported when the average score over the top three poses is used instead. The cluster population refers to the population of the corresponding cluster in the same column. The RMSD_{lig} refers to the average heavy-atom RMSD of the ligand calculated for the top cluster with the standard deviation in parentheses.

		$\mathbf{MD}_{apo}$	$\mathbf{MD}_{holo}$
sampling performance [%]	BGT-UDP	1.8	37.0
	RIC-NEO	23.0	45.5
	ABP-ALL	2.5	49.5
pose rank	BGT-UDP	7 (6)	1 (1)
	RIC-NEO	1 (1)	2 (1)
	ABP-ALL	91 (346)[b]	1(2)
cluster population	BGT-UDP	6	9
	RIC-NEO	80	152
	ABP-ALL	-	188
F_{nat}	BGT-UDP	0.72 ± 0.04 (0.75/4)	0.81 ± 0.07 (0.71/3)
	RIC-NEO	0.93 ± 0.00 (0.93/1)	0.90 ± 0.03 (0.87/4)
	ABP-ALL	0.47 (0.73)[b]	0.85 ±0.06 (0.93/3)
$RMSD_{lig}$ [Å]	BGT-UDP	2.0 ±0.3 (1.7/4)	0.7 ± 0.2 (0.54/3)
	RIC-NEO	0.9 ± 0.3 (0.57/1)	1.0 ± 0.3 (0.67/4)
	ABP-ALL	1.9 (1.5)[b]	0.9 ± 0.2 (0.59/3)

Table 3.13: Performance of HADDOCK in reproducing the experimental structures of the BGT-UDP, RIC-NEO, and ABP-ALL complexes in ensemble docking calculations for MD_{apo} and MD_{holo}.
[a]: Results correspond to the statistics of the top pose in clusters obtained using ligand interface RMSD clustering with a 1 Å cutoff and a minimum of four poses per cluster. The cluster rankings are based on the average score of the top four poses (the ranking based on the score of the top pose is reported in parentheses). The first values in the last two rows refer to the average statistics of the top four poses of a cluster in the semi-flexible refinement (it1 step) of HADDOCK, while values in parentheses refer to the statistics/rank of the best (smallest RMSD with respect to the reference) pose among the top four. The sampling performance was calculated as the percentage of poses within 2 Å from the experimental structure out of the 400 generated models (one docking run was performed from the ensemble of 500 MD conformations). The fraction of native contacts F_{nat} is defined as the number of native intermolecular contacts identified in a docking pose divided by the total number of contacts in the reference structure. The contacts are evaluated within a shell of 5 Å from the ligand in the experimental structures.
[b]: Since no acceptable clusters were obtained in this case, the reported statistics correspond to single-pose statistics for the first acceptable pose ($\leq$ 2 Å) and best pose (in parentheses).

		EDES$_{3w}$	**EDES$_{4w}$**
sampling performance [%]	BGT-UDP	15.8	20.5
	RIC-NEO	15.8	20.5
	ABP-ALL	21.3	13.8
pose rank	BGT-UDP	2 (3)	1 (6)
	RIC-NEO	1 (1)	1 (1)
	ABP-ALL	1 (1)	1 (5)
cluster population	BGT-UDP	33	20
	RIC-NEO	53	69
	ABP-ALL	20	5
F_{nat}	BGT-UDP	0.80± 0.02 (0.79/1)	0.72 ± 0.04 (0.71/3)
	RIC-NEO	0.90 ± 0.06 (1.00/3)	0.83 ± 0.06 (0.73/3)
	ABP-ALL	0.80 ± 0.00 (0.80/1)	0.72± 0.03 (0.73/2)
RMSD$_{lig}$ [Å]	BGT-UDP	0.8 ± 0.2 (0.50/1)	1.3 ± 0.4 (0.87/3)
	RIC-NEO	0.9 ± 0.3 (0.34/3)	0.9± 0.2 (0.62/3)
	ABP-ALL	0.8± 0.1 (0.74/1)	2.0± 0.1 (1.98/2)^c

Table 3.14: Performance of HADDOCK in reproducing the experimental structures of the BGT-UDP, RIC-NEO, and ABP-ALL complexes in ensemble docking calculations for EDES$_{3w}$ and EDES$_{4w}$. See caption of Figure 3.13 for details.
^{c}: It should be noted that a cluster was found ranking 10th with RMSD$_{lig}$ = 0.86 ± 0.12 and scoring as the top pose within standard deviations (-26.8 ± 0.4 vs -25.3 ± 1.1 a.u.)

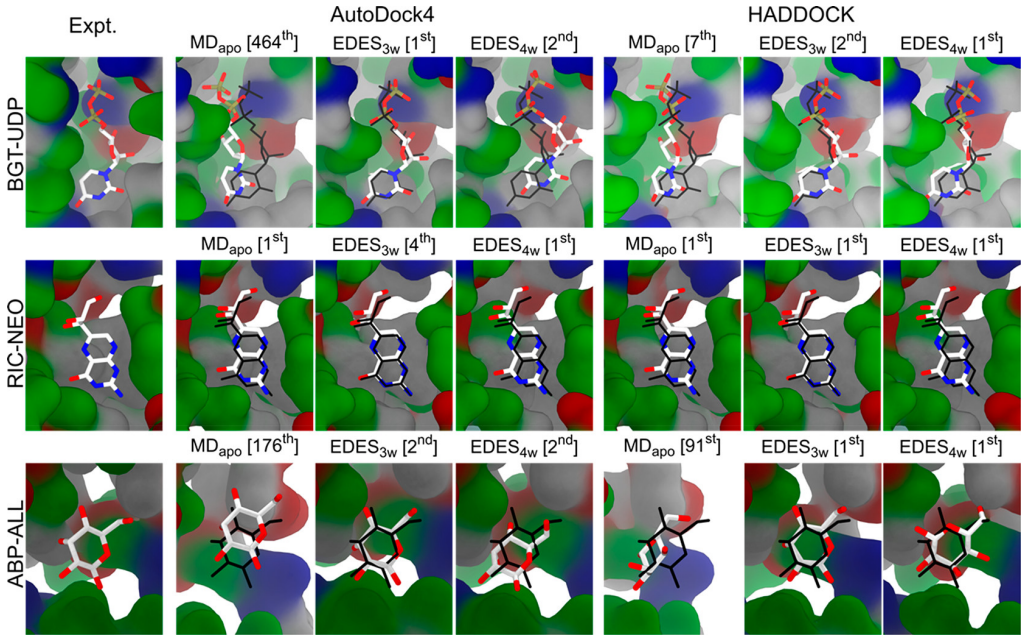

FIGURE 3.10: Docking performances of various structural ensembles in reproducing the experimental poses of BGT–UDP (top row), RIC–NEO (middle row), and ABP–ALL (bottom row). Close views of the experimental binding modes are shown in the first column, while the top-score poses within the first cluster with $\text{RMSD}_{lig} \leq 2$ Å, or poses with the lowest RMSD_{lig} value when no native-like pose was found, are reported in the next columns, with the corresponding ranks given in square brackets. The docking was performed both using AutoDock4 and HADDOCK for comparison. In each panel, the molecular surface of the backbone and that of the Cα atoms of the protein are colored by residue type as in Figure 3.1, and the ligand is shown as sticks colored by atom type. In columns 2 to 7, the experimental pose is shown in black thin sticks for easy comparison. Image taken from ref. [22].

3.1.5 Druggability assessment

We used the software f-pocket [236] to assess the druggability of the binding site for conformations generated by EDES (see chap. 2.6 for methodological details) and from the unbiased approach. The calculations have been limited to the cluster ensembles as they have been shown to well represent the whole conformational space sampled during the MDs. For each conformation, f-pocket gives a druggability estimation by means of a score (D) [237] ranging between 0 and 1, with higher values identifying more druggable geometries. It is customary to associate scores higher to 0.5 to putative binding sites [237]. Table 3.15 shows that EDES generates a much larger set of druggable structures than MD_{apo} for BGT and ABP. In particular, in the latter case no druggable conformation was generated by the unbiased simulation, while the performance of the two sets was similar for RIC, as expected. Moreover, in all cases EDES-derived ensembles have a higher percentage of structures associated with D higher than 0.5 than those derived from MD_{apo}. Finally, for BGT and RIC we can also see that the fraction of EDES-generated conformations featuring D higher than 0.9 is not much lower than that obtained from the MD_{holo}.

System		% of cluster structures with D greater than:				
Protein	Simulation	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
BGT	MD_{apo}	8.8	4.8	3.2	1.6	0.4
	MD_{holo}	40.0	28.6	18.0	10.6	2.4
	$EDES_{3w}$	15.8	10.2	6.8	4.6	1.4
	$EDES_{4w}$	14.2	10.8	7.2	3.6	1.8
RIC	MD_{apo}	2.6	2.2	1.2	0.4	0.2
	MD_{holo}	14.2	8.8	6.0	2.4	0.4
	$EDES_{3w}$	3.2	2.4	1.6	0.4	0.2
	$EDES_{4w}$	3.4	2.6	1.8	0.4	0.2
ABP	MD_{apo}	-	-	-	-	-
	MD_{holo}	20.4	13.0	7.2	3.6	1.6
	$EDES_{3w}$	7.6	5.2	3.0	1.8	0.4
	$EDES_{4w}$	7.4	5.4	2.6	1.2	0.2

Table 3.15: Performance of various MD simulations in generating druggable conformations of the binding site. For each protein and each simulation, the percentages of structures (over the 500 cluster representatives) featuring druggability scores D larger than 0.5 to 0.9 are reported in columns 3 to 7, respectively.

3.2 EDES in Cross-Docking calculations

3.2.1 Introduction

An accurate characterisation of receptor/ligand binding poses is of great importance for successful computer-aided drug discovery projects. For this reason, blind docking challenges, such as the Drug Design Data Resource (D3R) Grand Challenge (GC) [281–283], represent a excellent opportunities for developers to test and validate their rational drug design methodologies against experimental datasets. The D3R Grand Challenge is an annual event in which pharmaceutically relevant sets of protein-ligand complexes, for which the experimental structures have not been made publicly yet available, are selected for ligand pose predictions and binding affinity evaluations. This year, we decided to partecipate to the fourth iteration of the competition (GC4) to test our recently developed EDES approach, with minor modifications with respect to the original workflow [284]. In this edition, the selected target for the pose prediction stages is the beta-site amyloid precursor protein cleaving enzyme 1 (BACE-1), a beta-secretase 1 protein [285]. BACE-1 plays a crucial role in the early stages of in Alzheimer's disease, in which the β-amyloid peptides composing the amyloid plaques are generated [286–288]. Given its central role in the formation of β-amyloids, since long time BACE-1 enzyme has been recognised as a key target for developing therapies against the setting of Alzheimer's disease [289, 290]. The research interests for this enzyme is also testified by the great number of BACE-1 protein structures deposited in the Protein Data Bank (PDB) [36] at the beginning of the challenge (>300 on September 4th 2018), the majority of which are known inhibitors of this target. Here we report the performance of a hybrid approach for ensemble-docking [74, 78] that we developed for this challenge, coupling our recently proposed EDES protocol to sample holo-like and druggable protein conformations with the template-based algorithm for ligand conformer generation successfully

employed in the previous Grand Challenge 3 (GC3) competition [291]. With our workflow, near-native ligand poses were found for 16 (80%) and 17 (85%) of the twenty targets using, respectively, AutoDock [271] and HADDOCK [270, 280] as docking software. The most challenging cases were those for which the ligand generation steps produced conformers displaying the largest deviation from the geometry in the native complex, while virtually identical holo protein conformations have been obtained for all the cases. Importantly, our hybrid strategy performed best among the methods using receptor conformations generated without exploiting structural information of the enzyme bound to other ligands.

3.2.2 Methodological details

The D3R GC4 is composed of a set of different stages, in which participants were asked to predict the binding pose and/or calculate the binding affinity for a set of ligands when bound to a given protein receptor. The first two stages (hereafter stages *1a* and *1b*), in particular, aimed to (i) predict the crystallographic binding poses of twenty ligands into a receptor for which the bound conformation was not known (stage *1a*) and (ii) predict the binding poses of the ligands of stage *1a* but when the bound receptor conformations were made available (stage *1b*). So, for stage *1a* (a typical cross-docking experiment), we performed ensemble docking calculations of conformations of both the protein and the ligands generated by our methodology (vide infra), while in stage *1b* (an example of self-docking calculation), we used the same ligand conformations employed for stage *1a* but we docked them against the experimental bound conformations of the protein receptor, made available by the organisers for this stage. For both challenges, we submitted a set of 5 poses for each ligand, according to challenge rules.

3.2.2.1 Data provided

In stage *1a* the only data provided by the organisers consisted of a list of 20 compounds given in the *Simplified Molecular Input Line Entry System* (SMILES) code [292] together with the protein primary sequence given in the FASTA format [293]. In stage *1b* the experimental structures of the receptors for all 20 BACE-1 complexes were provided, to allow the participants to self-dock each ligand on the corresponding holo-conformation of the receptor.

3.2.2.2 Binding site determination

In our first work [22] (chap. 3.1), we used the experimental holo structure of each target to determine receptor's BS. In this case, as no holo structure was available, we followed the approach presented in ref. [202]. Namely, we retrieved in the PDB [36] all the protein structures featuring at least the 95% of sequence identity to the amino acid sequence provided by the organisers via the FASTA sequence and having a co-crystallised ligand. Among the entries, we discarded (undesirable) structures such as the ones having crystallisation buffer molecules as ligand(s), low resolution ones, structures with split side chains near the binding site and cases in which ligands were covalently bound, as we know ours bind non-covalently. This resulted in 340 structures. We verified that the binding site was well characterised and perfectly

conserved in all these cases, with no structure featuring missing residues in the putative pocket. Next, we used the Tanimoto metrics [294], as implemented in fmcsR [295] and chemmineR [296] packages to evaluate the similarity between the ligands present in the selected (340) structures and each of the 20 target compounds provided, in order to identify a set of receptor templates featuring the most similar ligand to the compounds to be docked.
If A and B are the two molecules to compare, we can define vectors $\vec{A}$ and $\vec{B}$ as the N-bit-long binary vectors encoding the fingerprints of the two molecules. Each bit of the two vectors represents a specific molecular feature chosen for the similarity calculation, and is set to "1" if the molecule posses that property and to "0" otherwise. Tanimoto coefficient is then defined as shown in eq. 3.1, where the quantities a and b represent the number of bits set to "1" for molecule A and B, respectively, while c is the number of common bits set to "1". Choosing the appropriate features to include in the Tanimoto similarity measurement, it serves as an accurate and easy way to compute the structural and chemical similarities between different molecules;

$$T_{A,B} = \frac{c}{a + b - c} \tag{3.1}$$

Further details on Tanimoto similarity measurement in the case of biological molecules can be found in refs. [202, 284, 291, 294]. The search for the structure featuring the most similar ligand to each of the 20 compounds resulted in 10 complex structures selected as templates (Table 3.16). From these structures, the residues lining the binding site were identified, following the same geometrical approach used in the previous work (chap. 3.1). To perform this task, a single list of residues was built by merging all the residues within 3.5 Å from the ligand in each of the 9 complex structures selected. With this approach, more than 30 residues were included in this preliminary definition of the BS. Among them, some residues were part of the putative binding region only in a single template structure while others were common to multiple templates. For this reason, we decided to consider (i) only the residues appearing at least in 2 structures (i.e. the most conserved ones) and (ii) among those present only in one structure, the most buried ones (likely to interfere with ligand binding). In this way the number of residues was decreased to 20. Figure 3.11 clearly highlights that the chosen list of 20 residues (Table 3.17) surrounds all the 20 congeneric ligands provided for this challenge.

Ligand target	Template PDB ID	Tanimoto similarity
BACE01	3DV1	0.605
BACE02	3DV1	0.875
BACE03	3DV1	0.821
BACE04	3DV1	0.872
BACE05	3DV1	0.725
BACE06	4DPI	0.660
BACE07	2IQG	0.618
BACE08	3DV5	0.543
BACE09	3DV5	0.698
BACE10	3K5C	0.739
BACE11	3VEU	0.833
BACE12	4KE1	0.681
BACE13	3K5C	0.681
BACE14	3K5C	0.861
BACE15	3K5C	0.891
BACE16	3K5C	0.750
BACE17	2B8L	0.490
BACE18	3DV5	0.476
BACE19	3DV1	0.625
BACE20	6BFD	0.604

Table 3.16: Ligand templates structures. For each target compound (first column), we report the template PDB ID (second column) and the TanimotoCombo similarity coefficient between the two ligands (third column), evaluated with the software software OpenEye ROCS. The TanimotoCombo coefficient ranges from 0 (lowest similarity) to 1 (highest similarity).

Residue	Occurrence
Q12	3
G13	1
D32	9
G34	9
S35	2
Y71	5
T72	9
Q73	7
G74	2
F108	4
F109	1
I110	1
Y198	1
I226	2
D228	9
S229	2
G230	9
V232	8
N233	1
T329	1

Table 3.17: List of residues defining the putative binding site of BACE-1 ligands investigated in this work. Residues are reported along with their occurrence frequencies (ranging from 0 to 10) in the list of residues within 3.5 Å of the ligands in the (experimental) template structures.

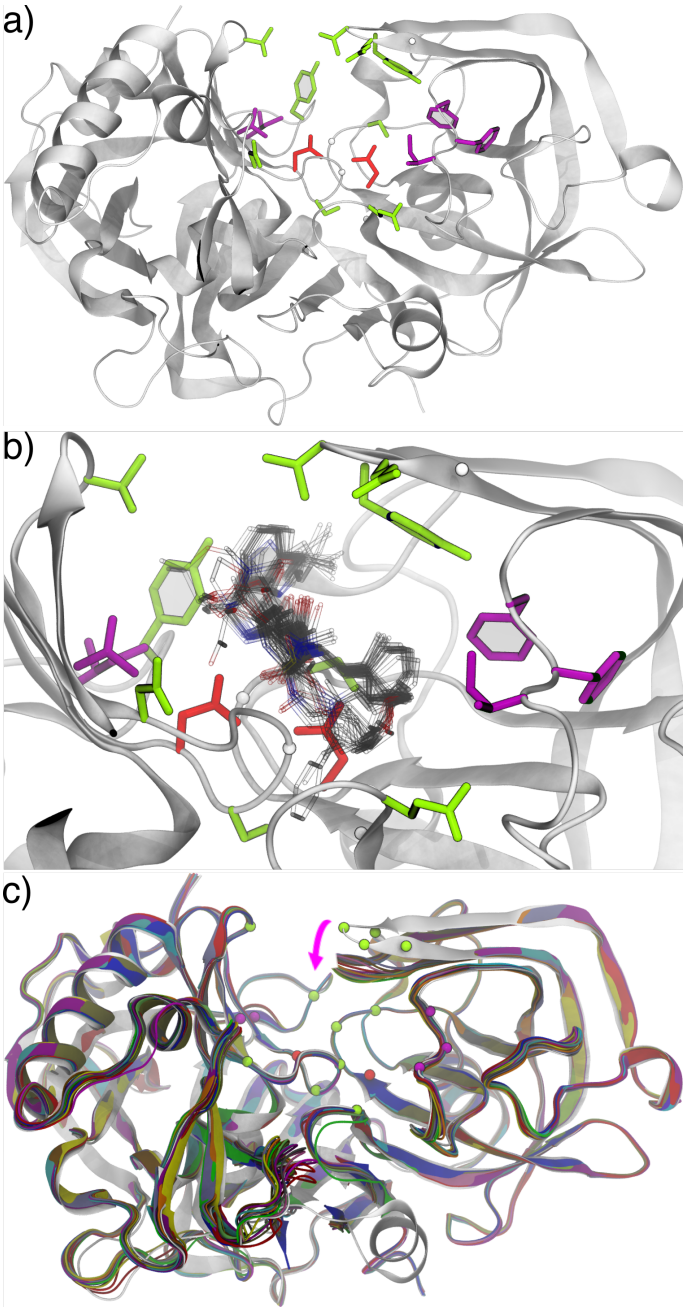

FIGURE 3.11: Putative binding site identified on the BACE-1 apo protein (PDB ID 1SGZ [297]) for implementation of the EDES approach. a) Structure of the protein (grey ribbons) showing the side chains of the 20 residues in Table 3.17 as sticks coloured by type (polar, apolar, acidic and glycines in light green, magenta, red and white respectively); b) zoom on the putative binding site in a), showing in transparent sticks the experimental poses of the 20 ligands provided by the organisers (after superposition of common Cα atoms on all proteins to 1SGZ); c) comparison between the apo structure of BACE-1 (1SGZ, grey ribbons) and the 20 ligand/BACE-1 complex structures (blue ribbons) released at stage *1b* of the challenge.

3.2.3 Ligand preparation

The 20 ligands were similar in size, each containing about 35 heavy atoms. Since ligands were provided in the SMILES string, a line notation for describing chemical structures, the first task was to generate their three-dimensional conformations. To do so, we employed the methodology featured in ref. [291].

As explained in the previous paragraph, the first step consisted in identifying a set of template receptor structures, highly homologous to receptor's sequence. In our case, this resulted in a set of 9 different template structures. Next we generated up to 500 conformers for each of the 20 ligand targets using the OpenEye OMEGA software package [298]. Then we calculated the TanimotoCombo coefficients (combining shape and chemical similarity) [291, 299] between all target conformers and template ones by means of the software OpenEye ROCS (using the "shape and colour" mode) [299] to select template structures with a similar holo conformation to the one assumed in presence of the ligand(s) to be docked. Finally, for each target compound, we selected the 10 conformers (among the 500 generated) displaying the highest Tanimoto-combo similarity to the corresponding template ligand.

3.2.4 Standard and enhanced-sampling (MD) simulations

In order to run the simulations, it is crucial to have a starting conformation of the protein. In this case, at odd with the previous work, no unbound (apo) conformation of the target was given by the organisers, as only the primary (1D) sequence of the receptor was provided.

In order to identify a suitable unbound target conformation, we searched the PDB database looking for protein structures displaying an homologous primary sequence to the one provided by the organisers. To do so, we used the software package BLASTP 2.7.1+ [300], setting the number of alignments to 1000 and the number of score evaluations to 10, while using the default values otherwise. We also searched only for entries having the keyword "BACE" in their name. With this approach, we identified around 300 structures, of which only 8 were apo proteins. Among these, we searched for conformations resolved at good resolution (2 Å or better), that did not feature any missing residue, in particular at the putative binding site, and displaying a full overlap with the BACE-1 sequence provided by the organisers. The only structure matching with these criteria was the one with PDB ID 1SGZ [297], which we selected as starting protein conformation for the simulations. However, prior to setting up the system, the structure was further refined through the MolProbity webserver [301].

3.2.4.1 Standard MD Simulation

Standard all-atom MD simulations of the apo protein (herafter MD_{apo}) embedded in a 0.15 KCl water solution ($\approx$ 60.000 atoms in total) and under periodic boundary conditions were carried out using the pmemd module of the AMBER18 package [136, 302]. The initial distance between the protein and the edge of the box was set to be at least 16 Å in each direction. Topology files were created for each system using the LEaP module of AmberTools18 starting from the apo structure with PDB ID 1SGZ. The AMBER-FB15 [43, 303] force field was used for the protein, the TIP3P-FB model was used for water, and the parameters for the ions were obtained

from ref. [304]. Long-range electrostatics was evaluated through the particle-mesh Ewald algorithm using a real-space cutoff of 12 Å and a grid spacing of 1 Å in each dimension. The van der Waals interactions were treated by a Lennard-Jones potential using a smooth cutoff (switching radius 10 Å, cutoff radius 12 Å). Multistep energy minimisation with a combination of the steepest-descent and conjugate-gradient methods was carried out to relax internal constraints of the systems by gradually releasing positional restraints. Following this, the system was heated from 0 to 310 K in 10 ns of constant-pressure heating (NPT) using the Langevin thermostat (collision frequency of 1 ps^{-1}) and the Berendsen barostat. After equilibration, a production run of 1 μs was performed. A time step of 2 fs was used for pre-production runs, while equilibrium MD simulations were carried out with a time step of 4 fs in the NPT ensemble (using a MC barostat) after hydrogen mass repartitioning [144]. Coordinates from production trajectory were saved every 100 ps.

3.2.4.2 Metadynamics Simulation

This step has been performed accordingly to what described in the general EDES workflow (chap. 2) and already discussed in the previous work. Here, only the main steps will be recalled, highlighting the small modifications with respect to the original recipe [284].

Metadynamics simulations were performed on the apo protein using the GROMACS 2016.5 package [277] and the PLUMED 2.3.5 plugin [278]. Simulations started from the last conformation sampled along the pre-production step of the unbiased MD. AMBER parameters were ported to GROMACS using the acpype parser [279]. We used four CVs, defined only on the binding site region: 1) the radius of gyration of the binding site (RoG_{BS}); 2-4) the numbers of (pseudo)contacts across three orthogonal "inertia planes" (CIPs). All non-hydrogenous atoms were considered to define the three CIPs, while only backbone atoms were used to estimate RoG_{BS}, at odd with the original approach in which also the latter CV was implemented considering all the heavy atoms. This change was made to reduce the computational cost of the calculation, considering that also the new implementation of RoG_{BS} is still able to act (although in an indirect way) on the BS' side-chains, which are however directly biased by the CIPs variables. Also in this case we applied the "windows" approach, aiming to sample, in a controlled manner, shapes of the binding site associated to decreased RoG_{BS} values. However, in this case, we only generated 3 windows including the first one, centred respectively at 9.91, 9.41 and 8.92 Å (corresponding to a global 10% RoG_{BS} decrease with respect to the one of the experimental apo structure, center of the first window). The choice to limit the extent of the conformational sampling only to 3 windows (instead of 4, as in the original approach) was manly due to the time restrictions of the challenge. In the first instance, this choice is supported by the findings that EDES sampling performance is not very sensitive to the usage of 3 or 4 windows (see chap. 3.1). Moreover, in the spirit of a further validation of the original protocol, also in this case we focus on the sampling of conformations displaying a partially collapsed binding sites (i.e. with smaller RoG_{BS} than the apo system). Although this choice can appear somehow arbitrary, as already discussed, it supported by several studies showing that ligand binding in enzymes most often results in a closed conformation of their

binding pockets as compared to the apo structures (see e.g. refs. [27, 305] or refer to the *Introduction*, chap. 1). Finally, for this specific system, the choice can also be justified a posteriori observing the (small) collapse of the binding site of BACE-1 occurring upon binding of all ligands (vide infra).
Each replica was simulated for 100 ns, leading to 400 ns of metadynamics simulations per window; coordinates were saved every 10 ps. The height w of the Gaussian hills was set to 0.6 kcal/mol, while the widths σ_i of the Gaussian hills were set to 0.06, 2.6, 1.7 and 3.0 respectively for RoG_{BS} and $CIP_{1,2,3}$. The bias factor for well-tempered metadynamics was set to 10. Hills were added every 2 ps, while the bias-exchange frequency was set to 20 ps. The force constants used "windows" approach were the same used in the previous work. Hereafter, the concatenated trajectory of the 3 windows will be refereed as $EDES_{3w}$.

3.2.4.3 Docking calculations

Docking calculations were performed on a set of receptor conformations extracted from the MD trajectories according to the clustering protocol presented in the *Computational Method* section (chap. 2) and discussed in the previous work (chap. 3.1). However, a few differences with the original implementation should be highlighted. First, the number of protein conformation clusters used for docking calculations here is reduced from 500 to 200. This was essentially due to the need to cope with the time constraints of the challenge, considering that here we used multiple conformations for each target ligand. Moreover, in the original EDES implementation, all the clusters were extracted from the metadynamics run. Docking performed on the clusters from the unbiased MD were indeed only used as a reference to compare the results of the two approaches. As an improved recipe of the original method, and in the spirit of developing a more general approach, here we also included conformations coming from an unbiased MD run. Indeed, as the extent of the conformational changes occurring at the binding site was not known a priori, including also low-energy states coming from small rearrangements of the unbound protein conformation appeared reasonable. In particular, in this case we expected to observe significant oscillations of the BS region close to the very flexible flap (Figure 3.11) also along MD_{apo}. We applied our original multi-step clustering protocol to both MD_{apo} and $EDES_{3w}$ trajectories, with the additional requirement to extract at least 10 cluster representatives from each of the 10 slices in which each RoG_{BS} distribution was binned, so as to include a certain number of structures also from poorly sampled regions. In this way, 500 clusters from each trajectory were extracted. Next, an additional cluster analysis using the same approach was performed on the pool of 1000 cluster representatives in order to generate the final ensemble of 200 structures. Finally, as already highlighted, 10 ligand conformations for each target ligand were selected and employed in ensemble-docking calculations. Calculations were performed using either AutoDock4.2 [271] or the HADDOCK2.2 webserver [270, 280], following the scheme presented in the first work (chap. 3.1).

In particular, with AutoDock, a docking campaign against all the 200 receptor conformations was performed with each selected ligand conformer, while in the case of HADDOCK, for each target ligand, the whole ensemble of conformers was submitted and used in a single docking run. However, both docking programs were used in two

different variants, generating the four sets of protein-ligand binding pose predictions that we submitted for the challenge. More in details, we used the standard AutoDock (receipt ID pe6zg) and HADDOCK (receipt ID kmtri) schemes, an AutoDock approach with a subsequent step of pose refinement and rescoring (herafter $Autodock_{rr}$; receipt ID nstab) and an approach in which HADDOCK was used including all hydrogen atoms of both binding partners (hereafter $HADDOCK_{all-Hs}$, receipt ID apue7). The first two approaches follow precisely the docking scheme presented in the first work (chap. 3.1), so they won't be discussed here.

The $Autodock_{rr}$ approach follows the *Autodock* procedure, with an additional step consisting in the relaxation of the top 10 docking poses by means of a multi-step structural relaxation performed with AMBER18 [302]. Systems were optimized in vacuum through three consecutive cycles of restrained structural relaxation (1000 cycles of steepest descent followed by up to 24000 cycles of conjugate gradients) followed by an unrestrained optimization (2000 cycles of steepest descent followed by up to 8000 cycles of conjugate gradients). During restrained relaxation harmonic forces of 0.3, 0.2, and 0.1 kcal·mol^{-1}· Å^{-1} (respectively for the first, second and third cycle) were applied on all non-hydrogenous atoms of the system. Long-range electrostatics was evaluated directly using a cutoff of 99 Å, as for the Lennard-Jones potential. The AMBER-FB15 [43, 303] force field was used for the protein, while the parameters of the ligands were derived from the GAFF force field [274] using the antechamber module of AmberTools18. In particular, bond-charge corrections (bcc) charges were assigned to ligand atoms following structural relaxation under the Austin Model 1 (AM1) approximation [306]. Note that, as for HADDOCK, the definition of a topology involving permanent bonding interactions allow for keeping the correct ligand cycle connectivity during refinement, while allowing some degree of flexibility such as changes in torsional angles and formation of H-bonds. Finally, the poses were rescored using the same scoring function of AutoDock employed to rank the original docking poses.

The $HADDOCK_{all-Hs}$ approach is same as the *HADDOCK* one, except for the inclusion of all hydrogens (and not only the polar ones) in the structures of the binding partners. Finally, in addition to the ensemble-docking calculations using receptor structures generated in silico for stage *1a*, we also performed self-docking calculations (only with the standard AutoDock approach) for the stage *1b* of the challenge. In this case, for each ligand, we docked its conformers employed in stage *1a* against the conformation of the receptor extracted from the corresponding holo experimental structure, released by the organisers at the beginning of stage *1b* (protocol $Autodock_{self}$, receipt ID qb4hg). All the remaining parameters were identical to those employed within the *Autodock* protocol.

3.2.5 Results

In this section we discuss the performance of our approaches at both stages of the challenge, together with highlighting possible drawbacks of the method and some directions for further developments. First, we discuss the docking performance of during the stage *1a*, in which we predicted the near-native ligand poses for a set of 20 ligands known to be BACE-1 binders. Evaluations of the predicted poses

were performed according to the data downloaded from the D3R website[3]. In particular, the accuracy of the poses was evaluated by calculating the RMSD of each ligand with respect to its experimental reference structure (considering only heavy atoms), after superposition of the binding interface areas. We performed a global evaluation in terms of the success rate of each approach on the whole set of 20 ligands together with a more detailed per ligand analysis. In this context, we also performed docking calculations of the ligand conformations generated for stage *1a* in the experimental unbound BACE-1 conformation to evaluate the need for bound-like receptor conformations to obtain near-native ligand poses. Then, we separately analyse the performances of the sampling of holo-like receptor conformations and of the generation of near-native ligand conformations.
Finally, we discuss the results obtained in stage *1b*, in which we self-docked our generated ligand conformations into the true bound receptor conformations, in order to evaluate the improvements obtained with respect to results of stage *1a* when the true receptor conformations are used.

3.2.5.1 Stage 1a challenge

The performance of the method is summarised in Table 3.18, in which the global performances of the different methods used in this work are evaluated for their ability to retrieve near-native poses for the 20 BACE-1 target ligands. In particular, in the following we'll refer to median and average RMSD values, calculated on the whole set of 20 ligands but considering, for each ligand, only the top pose or the nearest-native one. In the first case they'll be indicated as $RMSD^1_{med}$ and $\langle RMSD^1 \rangle$, while in the second one $RMSD^{min}_{med}$ and $\langle RMSD^{min} \rangle$, where the apices "*1*" and "*min*" indicate that the calculation is performed considering only the top pose or the one displaying the minimum RMSD (thus being the nearest-native one). Finally, we'll also consider the median and average RMSD values on all the five poses submitted for all each ligand. In this case they will be indicated as $RMSD_{med}$ and $\langle RMSD \rangle$ (Tables 3.19, 3.20). All these values are calculated over all the heavy atoms of the ligand, after the alignment of the protein binding interface region. With this metrics, the best results were obtained with the Autodock$_{rr}$ approach, by which the top 10 poses obtained with AutoDock have been further optimised and rescored. Table 3.18 shows that with this approach, we obtained $RMSD^1_{med}$ and $\langle RMSD^1 \rangle$ values lower than, respectively, 2 and 3 Å. Consistently, considering only the nearest-native poses, we obtained $RMSD^{min}_{med}$ and $\langle RMSD^{min} \rangle$ values lower than 1.5 and 2 Å. These results clearly show the positive impact of the refinement procedure with respect to the "standard" AutoDock protocol (Table 3.18). Considering the standard approaches, HADDOCK with standard settings performs very similarly to AutoDock, while an appreciable drop in the accuracy is caused by the explicit consideration of non-polar hydrogen atoms of the ligand during docking, in the HADDOCK$_{all-Hs}$ approach. Although a detailed explanation of this behaviour would require more systematic studies, which are out of the scope of this work, a reasonable explanation is that the change in ligands' volume, due to the inclusion of the non-polar hydrogens, have a dramatic impact on the ability of HADDOCK's search scheme to place the ligands within the buried (and rather small) binding site of BACE-1.

[3] https://drugdesigndata.org/about/grand-challenge-4-evaluation-results

	Averages			Median		
Protocol	$\langle RMSD^{min} \rangle$	$\langle RMSD^{1} \rangle$	$\langle RMSD \rangle$	$RMSD^{min}_{med}$	$RMSD^{1}_{med}$	$RMSD_{med}$
Autodock$_{rr}$	1.73±0.88	2.86±2.71	4.24±1.77	1.38	1.78	3.89
Autodock	2.48±1.82	3.10±2.57	4.41±2.10	2.07	2.25	4.28
HADDOCK	2.28±0.99	4.12±2.73	4.64±1.53	2.06	3.12	4.23
HADDOCK$_{all-Hs}$	3.19±2.26	4.83±3.50	5.96±2.18	2.66	3.10	5.76
Autodock$_{apo}$	3.78±2.94	5.67±3.72	5.17±3.34	2.47	3.51	3.49

Table 3.18: Overall performance of our protocols in retrieving near-native ligands conformations of BACE-1 ligands (rows 4 to 8) during stage *1a*. The performance of the Autodock$_{apo}$ protocol is also shown. The values are in Å.

Moreover, a global comparison in terms of the $\langle RMSD^{min} \rangle$ and $\langle RMSD^{1} \rangle$ metrics of the performances of all applicants is shown in figure 3.12. Within this metrics, our methods are placed in the middle-left and middle-right regions of the histogram plot (Figure 3.12), when compared to the results obtained by all the other applicants. However, an inspection of the details of the protocols used by the other participants (for those for which they were made available) and leading to better results than ours in terms of $\langle RMSD^{1} \rangle$ showed that our approach is the only one not exploiting experimental information on the bound receptor conformations. Indeed, at odd with those approaches, all our strategies are based on ensemble-docking calculations where receptor conformations are generated in silico from an experimental unbound receptor structure, exploiting no information on receptor structures complexed with similar ligands to the ones provided for the challenge.

Our performances can be also analysed from a different perspective, highlighting the accuracy of the different approaches when tested on each ligand separately. We report the results for each of the 20 BACE ligands in Tables 3.19 and 3.20 for the AutoDock and HADDOCK derived approaches, respectively. Moreover, the results are also represented in Figure 3.14. Globally, the tables show that with the above-mentioned approaches, we were able to retrieve at least one pose with $RMSD_{lig} < 2.5$ Å respectively in 15 (Autodock), 16 (Autodock$_{rr}$), 17 (HADDOCK), and 10 (HADDOCK$_{all-Hs}$) out of twenty cases, corresponding to success rates of 75%, 80%, 85%, and 50%. Moreover, the high standard deviations associated to the $\langle RMSD \rangle$ values for almost each ligand (Figure 3.14 and Tables 3.19, 3.20) was somewhat expected. Indeed our EDES approach displays the (desired) tendency to maximise the structural diversity of receptor conformations used for docking calculations, in particular at their putative binding site, resulting in a large diversity in the docking poses obtained. For this reason, evaluating our performances on the basis of the success rate appears much more of value than considering average RMSD values over all the poses submitted for each approach. In the following, we will focus our analysis only to the top three approaches: Autodock, Autodock$_{rr}$, and HADDOCK. First of all, tables 3.19 and 3.20 reveal that the most challenging ligand is BACE02, the only one for which we obtained poses featuring an $RMSD_{lig} > 3.5$ Å from the native conformation with all the approaches.

In particular for this target, the best performance is obtained with HADDOCK, by which the nearest-native pose features a $RMSD_{lig} = 3.9$ Å. Additional challenging ligands include BACE10, for which the best $RMSD_{lig}$ value is 2.8 Å (obtained with Autodock$_{rr}$), and to a minor extent BACE07, BACE09, BACE14, BACE16 and BACE18, for which 1 out of the three protocols was unable to find poses with $RMSD_{lig}$ values lower than 2.5 Å. Autodock$_{rr}$ demonstrated to be the top-performing method in terms of the success rate, reproducing at least one near-native pose among the top 5 for all ligands but BACE02, for which $\langle RMSD^{min} \rangle = 4.5$ Å. In principle, the issues with BACE02 could be amended by including also in the docking stage some degree of partner flexibility, although it has been shown by others that in such cases the (potential) improvement is reported to be system-dependent [2, 15]. In this case, however, a (small) improvement is actually noticeable in the results obtained using HADDOCK, which includes by default flexibility of both docking partners by means of short MD runs in the space of the torsional angles and by which a lower value of $\langle RMSD^{min} \rangle$ is obtained for this target. In this particular case, however, these results were in part expected. Evaluating the performances obtained in the separate steps of receptor/ligand sampling (Table 3.21) clearly shows that in this case, while being able to obtain virtually identical holo-like geometries, our workflow is less accurate in generating the native-like ligand conformations. Finally, we also compared our predictions with the typical scenario in which we generated ligands conformations but the only receptor structure available is the apo one.

We thus performed docking calculations with AutoDock using (i) the 10 ligand conformations used for the stage *1a* and (ii) the experimental apo structure of BACE-1 (PDB ID 1SGZ) that we selected as starting conformation for the simulations. Table 3.18 confirms that, as expected, this approach (Autodock$_{apo}$ in the table), shows a dramatic drop of the accuracy in retrieving near-native ligand poses with respect to the top performing three strategies of stage *1a*, due to the lack of inclusion of protein flexibility.

	Autodock			**Autodock$_{rr}$**		
Target ligand	RMSD^{min}	RMSD^{1}	$\langle RMSD \rangle$	RMSD^{min}	RMSD^{1}	$\langle RMSD \rangle$
BACE01	1.7	2.2	3.7±2.9	1.2	1.2	1.8±0.5
BACE02	**4.2**	4.2	6.5±2.7	**4.5**	4.5	7.9±2.9
BACE03	**2.8**	2.8	5.3±3.4	1.8	2.6	3.8±3.4
BACE04	1.5	1.5	2.9±1.3	1.1	1.1	2.2±1.0
BACE05	2.1	2.1	3.8±3.3	1.4	1.6	5.3±4.9
BACE06	1.1	1.1	1.7±0.7	1.5	1.6	2.1±0.8
BACE07	2.1	2.5	2.5±0.3	**2.6**	2.6	3.4±0.9
BACE08	1.2	1.7	3.3±4.0	1.0	1.0	4.9±5.0
BACE09	2.2	2.2	4.5±2.9	2.4	3.1	4.1±3.1
BACE10	**9.3**	9.4	9.4±0.1	**2.8**	10.0	8.2±3.0
BACE11	1.5	2.8	2.3±0.8	1.0	1.2	3.6±4.3
BACE12	1.4	1.5	1.8±0.5	1.2	10.3	6.8±4.7
BACE13	1.5	2.2	2.1±0.5	0.9	1.4	3.6±4.2
BACE14	**4.2**	10.8	8.2±3.6	1.3	2.5	4.9±3.3
BACE15	2.2	2.2	4.4±3.8	1.1	1.8	2.4±1.0
BACE16	2.5	2.5	4.2±3.6	**2.6**	2.6	4.4±2.9
BACE17	1.8	1.9	4.8±2.7	1.4	1.7	3.1±2.4
BACE18	**2.6**	2.6	5.3±2.5	1.8	1.8	4.7±2.5
BACE19	1.8	2.3	6.8±4.4	1.3	1.5	3.4±3.0
BACE20	2.1	2.1	4.9±3.4	1.6	8.5	4.0±3.0

Table 3.19: Summary of the docking results obtained with the Autodock-derived approaches for 20 BACE-1 ligands (data from https://drugdesigndata.org). All values are expressed in Å. RMSD^{min} values larger than 2.5 Å are bolded.

	HADDOCK			**HADDOCK$_{all-Hs}$**		
Target ligand	RMSD^{min}	RMSD^{1}	$\langle RMSD \rangle$	RMSD^{min}	RMSD^{1}	$\langle RMSD \rangle$
BACE01	1.5	2.0	2.8±1.7	**3.1**	3.1	4.7±2.7
BACE02	**3.9**	4.4	4.3±0.3	**10.2**	10.5	11.0±0.7
BACE03	2.5	3.5	4.2±2.7	**3.2**	3.5	5.7±3.2
BACE04	2.2	9.6	6.6±4.0	2.3	9.7	7.1±3.6
BACE05	1.4	3.1	3.8±3.1	**8.7**	9.8	10.0±0.9
BACE06	1.8	2.3	3.6±3.3	2.3	2.3	5.8±4.6
BACE07	2.4	2.6	4.5±3.2	**3.5**	3.6	6.2±3.5
BACE08	1.3	1.6	2.0±1.1	1.4	1.5	5.0±4.7
BACE09	**3.3**	10.0	8.4±3.0	2.2	2.9	5.3±3.8
BACE10	**5.5**	5.8	6.6±2.0	**3.5**	9.5	6.5±2.9
BACE11	2.0	4.5	3.7±1.0	1.7	1.7	5.4±3.9
BACE12	1.2	1.9	3.5±3.4	1.4	1.9	4.8±4.3
BACE13	1.5	1.5	3.9±3.2	1.8	1.8	2.2±0.4
BACE14	2.1	9.4	6.9±3.8	**2.9**	3.1	4.7±3.5
BACE15	2.2	3.0	3.7±1.6	**3.4**	9.4	8.5±2.9
BACE16	2.5	3.1	4.5±3.5	**3.0**	3.3	6.0±4.0
BACE17	1.9	5.7	5.2±1.9	1.7	2.3	3.0±1.9
BACE18	2.0	3.9	4.0±1.4	1.9	2.4	3.5±2.0
BACE19	2.3	2.3	5.1±3.3	2.4	9.6	8.1±3.2
BACE20	2.0	2.0	5.3±4.0	**3.2**	3.2	5.9±3.6

Table 3.20: Summary of the docking results obtained with the HADDOCK-derived approaches for 20 BACE-1 ligands (data from https://drugdesigndata.org). All values are expressed in Å. RMSD^{min} values larger than 2.5 Å are bolded.

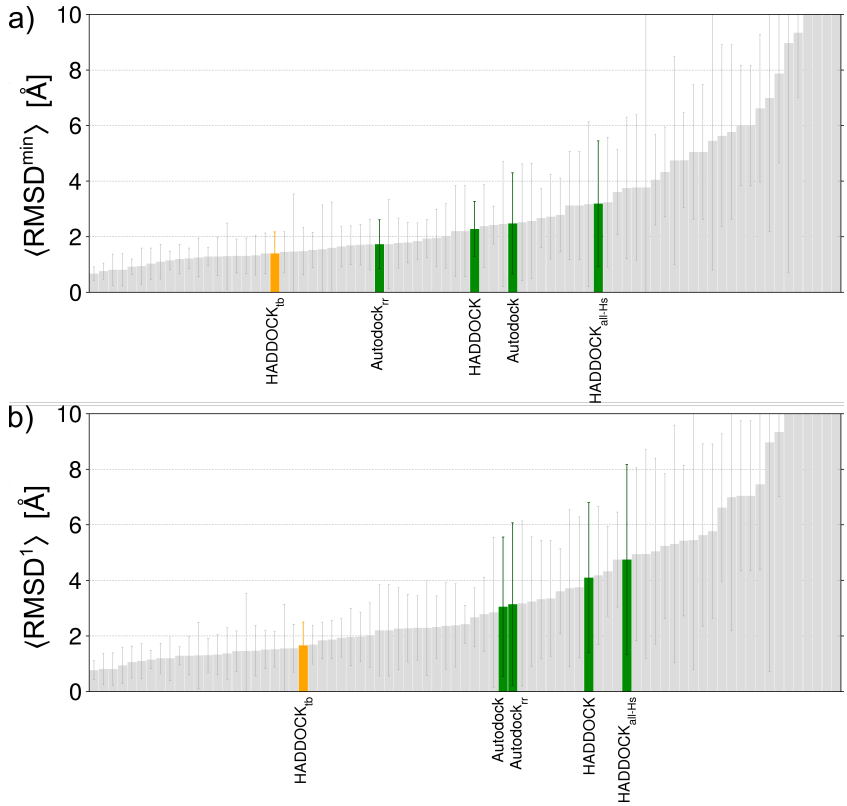

Figure 3.12: Overall performance of the docking protocols employed in this study, as measured by the values of $\langle RMSD^{min} \rangle$ and $\langle RMSD^{1} \rangle$

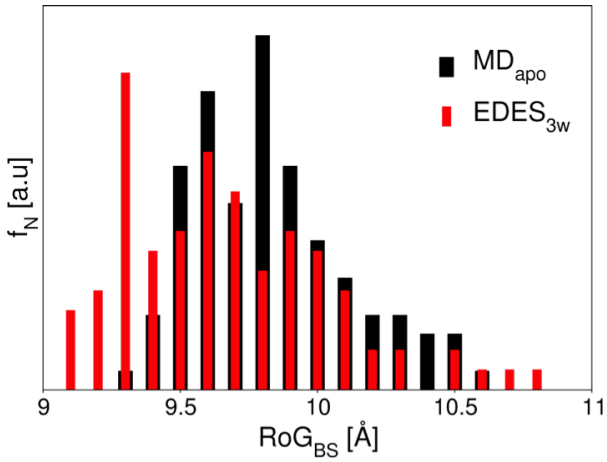

FIGURE 3.13: Distribution of RoG_{BS} calculated for the 200 conformational cluster representatives of BACE-1 selected for docking calculations. The clusters are divided in two groups according to the MD trajectory they were extracted from (103 and 97 clusters from MD_{apo} and EDES_{3w}, respectively).

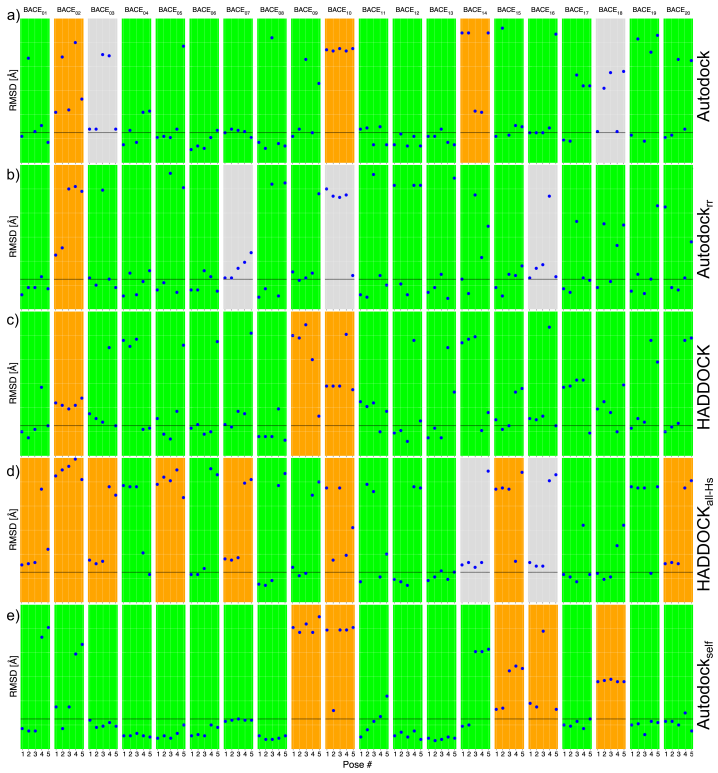

FIGURE 3.14: Performance of the Autodock (a), $\mathrm{Autodock}_{rr}$ (b), HADDOCK (c) and $\mathrm{HADDOCK}_{all-Hs}$ (d) protocols in reproducing the near-native conformations of the 20 BACE ligands. Green and grey panels refer to targets for which we obtained at least one pose within the top 5 featuring a value of the ligand RMSD $\leq$ 2.5 Å and $\leq$ 3 Å respectively, while orange boxes indicate cases for which no such poses were found among the top 5 ones.

3.2.6 Performance in sampling of holo-like conformations

In this section, we discuss about the ability of EDES to generate holo-like conformations of BACE-1 protein. This is evaluated in terms of the RMSD distributions calculated for the heavy atoms of the binding site (hereafter $RMSD_{BS}$) with respect to each of the 20 BACE-1 experimental structures (provided at stage *1b*). The analysis has been performed for the unbiased simulation of the apo system (MD_{apo}), for metadynamics one ($EDES_{3w}$) and for the ensemble of 200 cluster structures used in docking calculations. Results are shown in figure 3.15. Considering the structures with an $RMSD_{BS} \leq 2$ Å, the first evident feature is that both trajectories, MD_{apo} and $EDES_{3w}$, are able to generate a significant fraction of receptor conformations displaying a low $RMSD_{BS}$, with respect to every ligand/BACE-1 experimental structure. However, a careful inspection of the middle panel in figure 3.15 reveals the presence of a shoulder of low RMSD conformations more prominent in $EDES_{3w}$ than in MD_{apo}. This is also testified by the greater fraction of structures with $RMSD_{BS} < 1.5$ Å in the $EDES_{3w}$ run than in MD_{apo}. However, the overall good performance of MD_{apo} is not surprising, considering the (relatively small) conformational rearrangements undergone by the protein upon binding of all ligands (Figure 3.11). The small extent of structural rearrangements upon binding is also noticeable in the variation of RoG_{BS} in the apo/holo transition, specifically decreasing from the initial value of 9.79 Å (in the apo protein) to the range of 9.10-9.44 Å, found in the 20 holo conformations. As already pointed out, EDES method was primarily developed for targets undergoing large conformational changes upon binding, leading to a partial collapse the putative site. However, even its original recipe still allows to sample conformations with an enlarged pocket, with respect to the apo form, extending the possibility of usage of the method also for targets in which the pocket opens up upon binding. In the spirit of a further improvement in this sense, in the improved recipe used for this work we decided to include, in the ensemble of receptor conformations, also structures extracted from an unbiased MD simulation of the apo protein. Indeed, among the cluster representatives selected, a large diversity of RoG_{BS} has been observed, ranging from 9.1 Å to values over 10.8 Å. The largest fraction of structures featuring a collapsed pocket (smaller RoG_{BS} values) with respect to the unbound system (9.79 Å) derived from $EDES_{3w}$ as well as most structures displaying a RoG_{BS} close to the upper value (Figure 3.13). On the other hand, conformations coming from MD_{apo} displayed a RoG_{BS} distribution roughly centred around RoG_{BS}^{apo}. Moreover, together with a large variability of BS's compactness, our cluster analysis approach confirmed its ability to select a large (in some cases even larger than that sampled along the MD trajectories) fraction of low-$RMSD_{BS}$ geometries with respect to all the experimental reference structures (Figure 3.15 and Table 3.21). Finally, it is worth stressing that for all target receptors, we sampled at least a few conformations featuring the binding site virtually identical to the experimental structures, as testified by the lowest $RMSD_{BS}$ values of the clusters selected, all around 1 Å (Table 3.21).

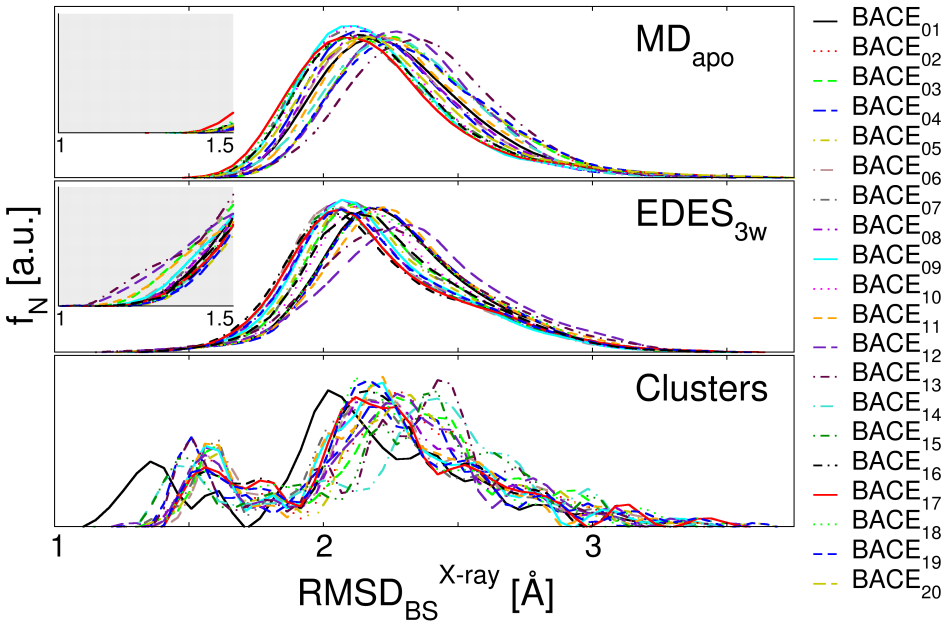

FIGURE 3.15: Normalised distributions (bin size = 0.1 Å) of RMSD_{BS} calculated with respect to the 20 experimental structures of ligands in complex with BACE-1. The analysis has been performed for MD_{apo} (upper panel) and EDES_{3w} (middle panel) trajectories, as well as for the ensemble of 200 BACE-1 structures used in ensemble docking calculations (lower panel). The insets in the upper and middle panels represent enlargements of the left-hand region of the corresponding graphs.

	Protein		**Ligand**
System	$RMSD_{BS}^{min}$ [Å]	% $RMSD_{BS} < 1.5$ Å	$RMSD_{lig-fit}$ [Å]
BACE01	1.13	16	1.11-1.71 (1.44±0.24)
BACE02	1.08	15	1.16-3.08 (2.70±0.63)*
BACE03	1.07	15	0.95-1.44 (1.25±0.15)
BACE04	1.08	19	1.03-2.05 (1.55±0.44)
BACE05	1.07	15	0.82-3.05 (1.50±0.73)
BACE06	1.19	15	0.58-1.38 (0.98±0.23)
BACE07	1.10	17	*1.62-2.77 (2.08±0.45)**
BACE08	1.07	17	0.57-1.51 (0.92±0.29)
BACE09	1.17	17	*1.52-2.38 (1.96±0.38)*
BACE10	1.13	17	*1.54-2.33 (1.98±0.33)*
BACE11	1.06	17	1.03-2.24 (1.64±0.48)
BACE12	1.07	17	0.68-1.79 (1.13±0.44)
BACE13	1.06	16	0.52-1.00 (0.75±0.16)
BACE14	0.98	16	1.37-3.34 (2.48±0.81)*
BACE15	1.16	14	1.53-2.23 (1.95±0.33)
BACE16	1.17	14	*1.83-3.53 (2.39±0.59)**
BACE17	1.13	14	1.08-1.58 (1.34±0.18)
BACE18	1.14	14	*1.53-2.01 (1.81±0.11)*
BACE19	1.07	16	1.17-1.51 (1.33±0.10)
BACE20	1.20	15	1.49-3.26 (1.83±0.52)

Table 3.21: Performances of our methodology evaluated separately for the generation of protein and ligand conformations similar to those found in the ligand/BACE-1 experimental structures. The 2nd column reports the lowest $RMSD_{BS}$ calculated across the 200 receptor conformations with respect to each experimental structure. The 3rd column reports the percentage of conformations displaying an $RMSD_{BS}$ lower than 1.5 Å. The last column reports the minimum and maximum RMSD values (calculated on the non-hydrogenous atoms with respect to the structure of each ligand in the experimental structure), as well as the average and standard deviation within parentheses. Values of $RMSD_{lig-fit}^{min}$ larger than 1.5 Å and average values of $RMSD_{lig-fit}$ larger than 2 Å are italicised and marked with an asterisk (*), respectively

3.2.7 Generation of near-native ligand conformers

In order to highlight the performance of our improved EDES recipe, in this section we discuss the ability of our template-based similarity protocol in generating near-native conformations for the 20 target ligands. Table 3.21 reports the statistics of the RMSD calculated on the heavy atoms of each ligand after it has been aligned on the reference experimental conformation (released in stage *1b* by the challenge organisers), hereafter refereed as $RMSD_{lig-fit}$. In all cases, the closest-to-native ligand conformation generated displayed a $RMSD_{lig-fit}$, lower than 2 Å, confirming the good sampling performance of this approach, as already reported in ref. [291], in generating at least one near-native conformation for all the different (macrocycle) ligands considered in this work. However, in general terms, the RMSD values obtained in ligand sampling are slightly larger than those obtained for the receptor (Table 3.21). In particular, for 6, 4 and 1 cases, corresponding to the 30, 20 and 0.5 %, we obtained an average $RMSD_{lig-fit}$ respectively greater than 1.5, 2 and 2.5 Å, while for the remaining cases the average $RMSD_{lig-fit}$ was smaller than 1 Å. Not surprisingly, the above-mentioned cases in which the average $RMSD_{lig-fit}$ is greater than 1 Å (Table 3.21), exception made for BACE03 and BACE15, are also the ones for which the docking results are the least accurate. To further investigate into the ligand generation strategy, figure 3.16 shows the histograms of $RMSD_{lig-fit}$ values for all (500) ligand conformers generated per each target together with the ones of the 10 selected for the docking runs by means of Tanimoto metrics. First of all, by the visual inspection of the figure, it is noticeable how the conformation generation protocol by means of the OpenEye OMEGA software is able to generate, in most cases, a large fraction of conformations similar to the native one ($RMSD_{lig-fit} <$ 2.0 Å). However, as we know that also a slight change in the orientation of a single ligand's functional group can have a dramatic impact on docking calculations, the lack, in all cases, of a considerable fraction of structures featuring a $RMSD_{lig-fit}$ $<$ 1.0 Å indicates the need to improve this step of the workflow. However, in the majority of cases, the selection criterion used (based on the Tanimoto similarity metrics) is able to include the ensemble of the 10 ligand conformations, a large fraction of low-RMSD geometries, a notable exception being represented by BACE02. Therefore, as already pinpointed by others [307–309], the generation of native-like conformations of macrocyclic ligands regards not only the selection process, but also (and even to a larger extent) the conformer generation step. Globally, the combined ligand generation/selection approach produced a large fraction of the 10 selected conformers with $RMSD_{lig-fit}$ values lower than 1.5 Å for 11 over the twenty targets (BACE01, BACE03, BACE04, BACE05, BACE06, BACE08, BACE11, BACE12, BACE13, BACE17 and BACE19), which are also roughly the ones for with we obtain the most successful docking results with the Autodock$_{rr}$ and HADDOCK protocols. On the other hand, ligand targets for which the ensemble of conformers featured the most distorted geometries (BACE02, BACE07, BACE10, BACE14, BACE15 and BACE16) were also the ones for which we obtained the worst docking performance.

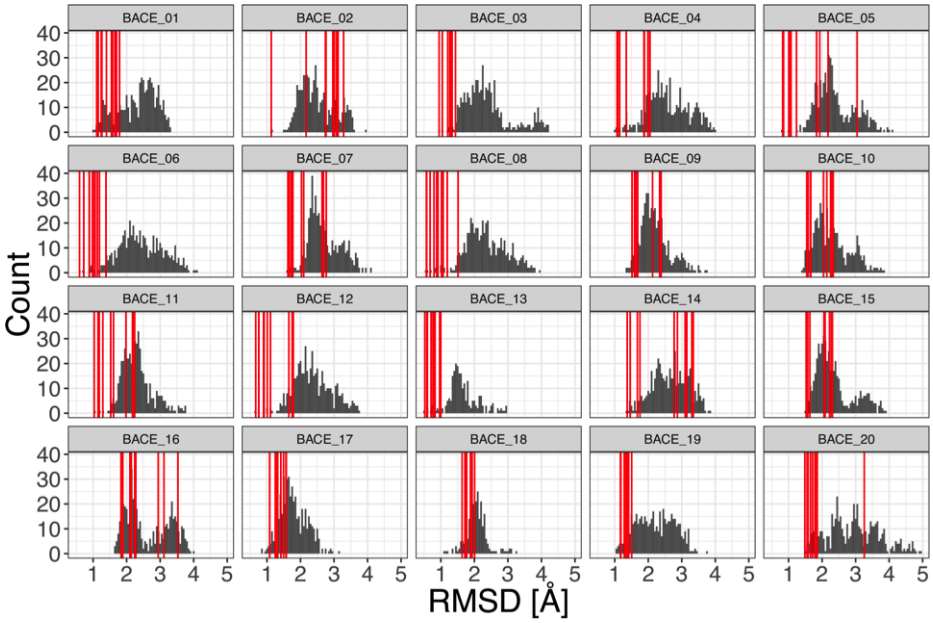

Figure 3.16: Histograms of RMSD values for all ligand conformers generated per target. The x axis shows the heavy-atom RMSD (Å) of every conformer after optimal superimposition on the crystallographic compound. The red lines highlight the RMSD values of the 10 conformers that were selected for docking based on the shape matching procedure outlined in the thesis.

3.2.7.1 Stage 1b challenge

Finally, in the following we'll discuss the results obtained in this self-docking stage in which we docked the 10 ligand conformers used for stage *1a* into the experimental bound protein structures (made available by the organisers for this stage). This step thus helps to highlight the accuracy in the generation of ligand conformers and of possible limitations of the workflow not directly linked to the treatment of protein flexibility. We performed this exercise only using AutoDock, using the same scheme followed during stage *1a*. In terms of averages (over the 5 poses submitted) we obtained 2.24±2.13, 2.93±2.78 and 3.59±2.73 Å respectively for $\langle RMSD^{min} \rangle$, $\langle RMSD^{1} \rangle$ and $\langle RMSD \rangle$. In terms of median values, on the other hand, we obtained 1.60, 2.03 and 2.30 Å for $RMSD^{min}_{med}$, $RMSD^{1}_{med}$ and $RMSD_{med}$. Results are also shown in Figure 3.14-e. Interestingly, the usage of the true holo receptor conformations results only in a marginal improvement with respect to the results of same protocol used at stage *1a* (Table 3.18), while it did not affect the success rate evaluated in terms of number of ligands featuring at least one pose with $RMSD_{lig-fit} \leq 2.5$ Å, which remained the 75%. This result is not surprising, in virtue of the good sampling of holo-like receptor conformations obtained with the (improved) EDES recipe. Moreover, we also note (Figure 3.14-e) that ligands shown to be the most challenging ones in stage *1a* still represent the ones for which reproducing the correct binding geometry is most difficult. In particular, no native-like poses ($RMSD_{lig-fit} \leq 2.5$ Å) were found for BACE09, BACE10, BACE15, BACE16, and BACE18, while for BACE07 all poses have RMSD values close to 2.5

Å, following the same trend observed in stage *1a*. While, in such cases, structural relaxation of the poses, followed by a rescoring step, as done e.g. in the Autodock$_{rr}$ approach, is expected to improve the results, we can note that, for cases like BACE02 and BACE14, also a very minor structural rearrangement towards the correct BS geometry was sufficient to find at least one native-like pose (Figure 3.14-e).

3.2.8 Conclusions

In this work, we reported the performance of our hybrid docking approach in its participation to the D3R Grand Challenge 4 competition [284]. Our workflow, based on ensemble-docking calculations, involves a template-based approach to generate and select a pool of ligand conformers coupled our EDES protocol (implemented for this work with small modifications with respect to the original version) to sample and select holo-like protein conformations starting from the apo one. Regarding the generation of ligand conformers, a good accuracy in generating and selecting near-native structures has been observed for most ligands, while EDES method confirmed its great performance in sampling holo-like BACE-1 geometries for all congeneric target ligands. In particular, the results obtained in the case of BACE-1, undergoing only minor movements of a flap region upon binding, strengthen the possibility to use EDES protocol also for targets involving small structural rearrangements upon binding. These findings are reflected in the overall relatively good performance obtained in stage *1a*. Regardless of the specific approach used, we were able to find near-native poses among the top 5 ones for at least 75% of the twenty complexes subject of the pose prediction sub-challenge (stage *1a*). More in details, while HADDOCK was able to find near-native poses for more targets than AutoDock, the latter, coupled to a computationally cheap post-docking relaxation and rescoring of the poses, displayed the best overall performance among the four approaches presented. Finally, we also performed docking calculations for all target ligands on the experimental apo and holo conformations. In the first case, docking results proved to be significantly less accurate than those obtained with the EDES-obtained holo-like conformations, due to the incorrect positioning of the flexible flap region in the apo structure. On the other hand, performing docking calculations on the experimentally-obtained holo conformations produced only overall small pose-prediction improvements with respect to using EDES-obtained geometries. This confirms the ability of EDES approach to generate conformations prone to correctly host the ligand(s) and pinpoints to its general applicability, although originally developed only for targets undergoing large conformational changes upon binding.

3.3 Tackling very challenging systems: adenylate kinase

As third application of the EDES workflow, here we present some preliminary results concerning the sampling of holo-like structures for a further protein, the adenylate kinase, featuring (i) very large conformational changes upon binding, also involving secondary structure rearrangements and (ii) a very extended binding region, composed of two (sub)pockets with different physico-chemical properties. Namely, the two binding regions are known to bind different ligands via an allosteric behaviour, where the binding of the first substrate in its binding pocket increases, via extended conformational changes, the affinity for the ligand of the other binding pocket [310–313].

Adenylate kinase (hereafter AK, PDB ID apo: 4ake [314], PDB ID holo: 1ake [315]) belongs to the category of phosphotransferase enzymes and its role is central in the metabolism of adenine nucleotides. Moreover, its deficiency and/or malfunctioning has also a role in the onsetting of different pathologies, such as the haemolytic anemia [316]. Over the years, several works [25, 81–83] addressed the issue of reproducing holo-like geometries of this target from its unbound structure, showing that AK represents an extremely challenging system.

To address this target with our approach, we first identified the binding region from the experimental holo structure, considering the residues within 3.5 Å from the ligand, bis(adenosine)-5'-pentaphosphate (AP5), bound in the complex (PDB ID 1ake). The residues lining the binding site (hereafter BS_{exp}) are reported in table 3.22, from which it can be seen that the site is rather extended, containing 30 residues. We also checked, using this definition of the BS, the RMSD considering all heavy atoms and only backbone ones between the apo and holo conformations, obtaining respectively 5.1 and 4.8 Å. The extent of the conformational changes is also reflected in the variation of the RoG in the apo/holo transition, going from 14.5 to 11.2 Å and from 14.2 to 11.4 Å when calculated, respectively, over the heavy atoms and only over backbone ones. On the other hand, the partial collapse of the binding site due to ligand binding can also be noticed in a 23% RoG decrease in the first case and of the 20% in the second one. This makes AK a perfect challenging test case for our improved recipe not exploiting any experimental information on the holo conformation, not even for the BS identification. For this reason, we put ourselves in the condition of not having available the holo conformation and identified the BS region to be used for the simulations by means of the site-finder software COACH-D [257]. The software outputs a set of 10 possible binding sites, together with a rank (C) reflecting the confidence of the prediction. The score ranges from 0 to 1, where a higher score indicates a more reliable prediction. We identified a consensus binding competent region by merging the residues belonging to the first three identified binding sites (BS_{tot}), having a C-score respectively of 0.99 (site-1), 0.89 (site-2) and 0.67 (site-3). The remaining predictions all presented a C-score of 0.1 and for this reason have been discarded. Interestingly, site-1, site-2 and site-3 contained respectively 13, 33 and 18 and residues, with site-1 and site-3 being almost complementary to each other to generate site-2. Moreover, site-2 encoded 29 over the 30 residues of BS_{exp}. Next, we followed the same approach described in the other works presented here to perform an unbiased MD simulation (10 μs long) and a set of biased (metadynamics) simulations, following the same protocols presented in

the other works. However, in the case of the biased simulations, together with the approaches presented so far, we developed a new one based on an improved set of CVs. Several works [18, 20, 27, 31, 305, 317] highlighted that a common and widespread class of rearrangements in the context of ligand binding is represented by domain hinge-movements, which can be interpreted as a (quasi-rigid) rotation of one domain of the protein relative to another. In the spirit of being able to enhance this specific class of motions together with the unspecific structural rearrangements triggered by the usage of CIP variables, we introduced a new class of collective variables, hereafter simply refereed as "contacts between rigid domains" (CRD). Defined in the same way of the CIPs, these CVs specifically target hinge-like motions between quasi-rigid domains. First, an analysis on the structural dynamics of AK has been performed by means of the software SPECTRUS [318], identifying three quasi-rigid domains: a main (central) one (CORE), surrounded by two smaller ones placed on two opposite sides, the ATP-binding domain (LID) and the NMP-binding one (NMP) (Figure 3.17). The binding site region extends through all the CORE domain up to its borders with LID and NMP (reflecting the (sub)pockets identified COACH-D as site-1 and site-3).

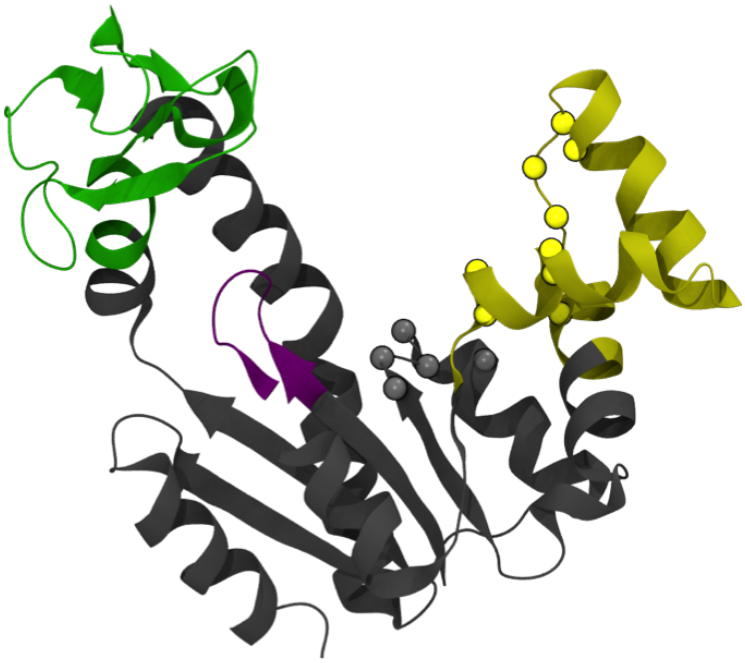

FIGURE 3.17: Adenylate kinase (AK) apo conformation, with the three (quasi)-rigid domains identified by the software SPECTRUS [318] highlighted in different colors: black for the central (CORE) domain and green and yellow respectively for the LID and NMP ones. Residues of the extended binding region are also represented as spheres corresponding to their centres of mass. Dark purple, green and yellow spheres indicate respectively the residues in the CORE, LID and NMP domain.

Then, we divided the binding region into two sets of facing residues, respectively at the CORE/LID and CORE/NMP interfaces. To identify the residues of the binding region at both interfaces we selected BS_{tot} residues within a 8 Å cutoff from any residue belonging respectively to the LID and NMP domain. This identified

respectively the two regions refereed hereafter as BS_{LC} (LID/CORE) and BS_{NC} (NMP/CORE) (Table 3.22). The CDRs were then defined according to the CIP scheme, considering all the heavy atoms of the corresponding selections.

In the following, we'll present the results in terms of sampling of holo-like conformations of this target for (i) a standard unbiased MD simulation (MD_{apo}), (ii) the standard EDES approach with 4 windows ($EDES_{std}$) and (iii) an enhanced-sampling approach combining the standard RoG/CIP set of CVs to the new CRD ones, with a total 6 CVs ($EDES_{crd}$). In the latter case, both RoG and the three CVs have been defined on the total binding site BS_{tot}, while the CRDs have been defined on BS_{LC} and BS_{NC}. Moreover, in this case, together with the already mentioned difference in the set of CVs used, the windows approach has been converted into one only employing a variable upper restraint where the RoG value of BS_{tot} is used to gently drive the sampling into structures with a decreased RoG value. Simulations start with a very soft restraint on RoG_{apo}^{X-ray} with a force constant of 10 kcal·mol^{-1}· Å^{-1} for 1 ns, which is increased of the same amount each ns, until the force constant value of 50 kcal·mol^{-1}· Å^{-1} is reached. Then, each 10 ns the value at which the restraint is set is decreased of a fixed amount, in order to push the systems towards conformations featuring a value of RoG decreased of the 20 % with respect to RoG_{apo}^{X-ray}. Each of the 6 replicas is 500 ns long, for a total simulation time of 3 μs. The results obtained with these approaches will also be compared to the ones of other works addressing the same target (Table 3.23).

Protein	**Binding site's residues**
BS_{exp}	A8,P9,G10,A11,G12,K13,G14,T15,T31,G32,L35,R36,M53,L57,L58 V59,V64,N84,G85,F86,R88,Q92,R119,R123,D158,R167,K200,P201,V202,V205
BS_{LC}	A8,P9,G10,A11,G12,K13,T15,R119,R123,D158,R167,K200,P201, V202,V205
BS_{NC}	T31,G32,L35,R36,M53,L57,L58,V59,V64,N84,G85,F86,R88,Q92

Table 3.22: BS definition for AK. The table reports the residues identified from the experimental holo conformation (BS_{exp}) as well as the two binding regions (BS_{LC} and BS_{NC}) identified by the site-finder software COACH-D [257].

		Approaches		
Protein region	**Results**	**MD_{apo}**	**$EDES_{std}$**	**$EDES_{crd}$**
BS_{LC} (noH)	Minimum RMSD (Å)	1.43	1.31	1.41
	% RMSD $<$ 2.5 Å	10.9	32.2	25.6
BS_{NC} (noH)	Minimum RMSD (Å)	1.66	1.48	1.54
	% RMSD $<$ 2.5 Å	2.9	3.2	1.2
BS_{exp} (noH)	Minimum RMSD (Å)	3.4	2.6	2.2
	% RMSD $<$ 2.5 Å	0	0	0.01
Protein (BB)	Minimum RMSD (Å)	3.7	2.4	2.2
	% RMSD $<$ 2.5 Å	0	0.002	0.1

Table 3.23: Performance of an unbiased MD simulation together with two enhanced-sampling approaches in sampling holo-like conformations for the AK protein. For each approach, the RMSD of the nearest-holo geometry is reported (minimum RMSD), together with the percentage of structures with an RMSD lower than 2.5 Å. The calculations have been performed on all heavy atoms for both the identified binding regions LID(NMP)/CORE and for the experimental binding region while the analysis has been restricted to only the backbone atoms for the whole protein.

Although the results presented in table 3.23 are still very preliminary and the work is still on-going, a few considerations can be already drawn. First of all, from the data presented it appears that not even a 10 μs long unbiased MD simulation (MD_{apo}) is able to sample holo-like conformations of such a complex binding region. Indeed, the minimum value of BS_{exp} is 3.4 Å. However, the unbiased approach is able to reproduce the individual binding regions BS_{LC} and BS_{NC} rather well, respectively with minimum RMSDs of 1.43 and 1.66 Å (from the values in the experimental structures respectively of 2.3 and 2.5 Å), even if the rearrangements towards the holo geometry of the whole binding site (BS_{exp}) fail to combine well. On the other hand, considering the biased approaches, we see that in terms of minimum RMSD both of them improve the sampling performance compared to MD_{apo} for the individual binding regions and for BS_{exp}. However, $EDES_{crd}$ performs better than $EDES_{std}$ when reproducing BS_{exp}, for which it is the only approach to generate conformations featuring a RMSD $<$ 2.5 Å. This might be linked to introduction of the new CVs explicitly targeting hinge-like motions, displayed in this target when binding its ligand AP5. However, it should also be noted that in this case the windows approach of the original EDES method has been converted into the usage of only variable upper restraint, so for a fair comparison the CVs used for the $EDES_{crd}$ approach should also be used within the original EDES protocol. Clearly further studies are needed in this sense. Finally, concerning the ability of the protocol to drag the whole protein towards holo-like conformations, we see the same trend observed in reproducing BS_{exp}. In particular, concerning $RMSD_{protein}$, our results can be compared to those of Kurkcuoglu and Doruker [25] and Ahmed et al. [317]. Using different sampling schemes, the nearest-to-holo conformation generated in both cases displayed a $RMSD_{protein}$ (considering the backbone atoms) of 2.4 Å.

Similar results are obtained also with our enhanced-sampling approaches, while such a comparison is unfeasible for the binding site, since details of the reproduction of this region were not disclosed in the other works. Docking calculations on the clusters extracted from these trajectories should be now performed to assess the quality of

those geometries in the reproduction of near-native ligand conformations. Moreover, other combinations of CVs and sampling protocols could also be considered, to further improve the understanding of the dynamical behaviour of this target. Finally, it should be stressed again that in this case no experimental information on the holo state has been used, not even to define the binding region, further confirming the possibility to use such approaches also when the binding region is defined by means of site-finder software.

Conclusions and future perspectives 4

4.0.1 General considerations

In this thesis we have presented a novel protocol able to generate holo-like and druggable protein conformations starting only from the knowledge of the apo structure. The method employs metadynamics on a new set of CVs, specifically targeting the putative binding site and resulting in the generation of maximally diverse protein geometries, including a relevant fraction of holo-like and druggable ones. Moreover, the usage of an ad-hoc designed clustering protocol allowed us to select a tractable (small) number of conformations while maintaining or even increasing (compared with the distributions obtained from the MD simulations) the fraction of holo-like geometries. We tested this protocol in the framework of ensemble-docking, performing both re-docking and cross-docking calculations. In the first case, the experimental ligand geometry was docked into the receptor conformations generated with our approach starting from an available X-ray unbound structure. We performed this exercise with three different targets, undergoing different extent of conformational changes upon binding and which had already proved to be challenging proteins for docking calculations. In the second case, we tested the general applicability of the whole workflow in D3R Grand Challenge 4 (GC4). Aim of the challenge was to retrieve near-native ligand poses for a set of 20 (macrocycle) ligands known binders of the BACE-1 enzyme. In this case, however, neither the 3D structures of ligands nor of the receptor were made available by the organisers. Finally, we also reported the preliminary results of applying a modified EDES protocol for the generation of holo-like conformations of the adelynate kinase, a protein featuring an extended binding region and undergoing the largest conformational changes among the other targets addressed here.

Given the very encouraging results obtained in all cases, we are confident that our protocol could pave the way towards an automated workflow for the generation of holo-like and druggable conformations of proteins, still today representing a limiting factor in structure-based drug design approaches.

4.0.2 Perspectives

A straightforward way to improve the sampling could be to couple the MD simulations with the use of co-solvents [61, 319] as done for example in ref. [320]. Furthermore, our original set of CVs could be improved by explicitly including other orthogonal degrees of freedom, such as global protein motions [25, 321, 322], rotations around torsional angles [24, 65] or secondary structure changes [167, 323]. Alternative routes to post-process docking poses can also be used, such as pose refinements with re-scaled protein-ligand interactions in explicit solvent [164, 320]. Moreover, experimental information from various sources on both the apo and holo states as well as bioinformatics predictions could be used at any stage of the process. Examples include the encoding of new ad-hoc CVs able to enhance the sampling along specific (known to be crucial) motion modes or the usage of restraints in both the conformational sampling/selection steps. With this respect, for example, experimental data on the gyration radius of the complex (often experimentally easier to obtain than the whole characterisation of the complex [324]) could be used as restraints to drive the sampling via the already implemented "windows approach" as well as used as filters during the clustering procedure [25].

Concerning the identification of the putative binding region, an important improvement would be the setting up of an automatic protocol for binding site(s) identification using consensus data from multiple site-detection programs [257, 325–328].

Our long term perspective would be to set up a database of protein conformations, in which, for each target, a number of maximally diverse and druggable protein geometries is provided. Moreover, the provided conformations should not be biased towards specific ligand chemotypes, so that they could be used also in the assessment of the binding properties of new therapeutic agents.

Acknowledgements

Three years have passed by, since when I made the decision to start this PhD and so it's time to officially thank all the people who have helped me during this path.

Firstly, I would like to thank my supervisors Paolo and Attilio for keeping me under their wings and let me be part of their research group. They gave me the amazing opportunity of pursuing a PhD and helped me during all the way to the end. They really gave me unreserved support (considering my quite odd personality :D) and helped me both academically and personally, right from the first day of my PhD (and also before, since I also graduated with them!). I can surely say these three years have been the most exciting ones I've ever had. So thank you again for your guidance and for the passion for Science that you all shared with me! Another great person I feel I need to say thanks to is Giuliano. I am really grateful to him for the nice discussions we had and for all the help and support he gave me during these years. His nice words, said with a smile, changed more than once the flow of my day. My deep and sincere thanks also to Prof. Alexandre Bonvin for giving me the opportunity to join his group as a visiting PhD student in the Netherlands. Thanks for the opportunity, for your kindness and for your help. The experience will certainly never be forgotten. Moreover, I would also like to thank Dr. Rudolf Oldeman and Dr. Carlo Carbonaro for their productive comments and stimulating questions during the yearly assessments of my scientific progress within the PhD program. Also, a big thanks goes to Dr. Alessandro Pandini and to Prof. Martin Zacharias for the useful suggestions they gave me in order to improve this manuscript. I am also very thankful to Andrea and Giovanni for their technical support and patience, and for all the software that I asked them to install!

Thanks also to all the present and past members of Paolo's group, such as Alessio, Chiara, Ivana, Francesco, Venkata and Francesca, for the time spent together, the support and all the (academic and not) discussions we had during these years. Finally, I would also like to thank my family, all the friends who helped and supported me during these years and somebody special who has joined my life since the 21st of august 2018. You all have been on my side during this experience and we lived together gloomy and happy times. Some of you also came to visit me while I was in the Netherlands. Thanks a lot for your time and for all the love and support you showed me.

I also gratefully acknowledge the Sardinia Regional Government for the financial support of my Ph.D. scholarship (P.O.R. Sardegna F.SE., Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2014–2020—Axis III Education and Training, Thematic Goal 10, Priority of Investment 10ii, Specific Goal 10.5., Action Partnership Agreement 10.5.12).

Bibliography

[1] Riccardo Baron and J. Andrew McCammon. Molecular recognition and ligand association. *Ann. Rev. Phys. Chem.*, 64(1):151-175, 2013.

[2] Du X, Li Y, Xia YL, Ai SM, Liang J, Sang P, Ji XL and Liu SQ. Insights into protein-ligand interactions: Mechanisms, models, and methods. *Int. J. Mol. Sci.*, 17(2): 144, 2016.

[3] Salahudeen MS and Nishtala PS. An overview of pharmacodynamic modelling, ligand-binding approach and its application in clinical practice. *Saudi Pharm. J.*, 25(2): 165-175, 2017.

[4] Hulme EC and Trevethick MA. Ligand binding assays at equilibrium: validation and interpretation. *Br. J. Pharmacol.*, 161(6):1219-37, 2010.

[5] Singh J, Petter RC, Baillie TA and Whitty A. The resurgence of covalent drugs. *Nat. Rev. Drug Discovery*, 10(4): 307-17, 2011.

[6] Tummino PJ and Copeland RA. Residence time of receptor-ligand complexes and its effect on biological function. *Biochemistry*, 47(20):5481-92, 2008.

[7] Cozzini P, Kellogg GE, Spyrakis F, Abraham DJ, Costantino G et al. Target flexibility: an emerging consideration in drug discovery and design. *J Med Chem*, 51(20):6237-55, 2008.

[8] D. D. Boehr, R. Nussinov, and P. E. Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.*, 5(11):789–796, 2009.

[9] R. B. Fenwick, S. Esteban-Martín, and X. Salvatella. Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *Euro Biophys J*, 40(12):1339–1355, 2011.

[10] A. D. Vogt, N. Pozzi, Z. Chen, E. Di Cera. Essential role of conformational selection in ligand binding. *J Biophys Chem*, 186(1):13-21, 2014.

[11] J.-P. Changeux and S. Edelstein. Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biology Reports*, 3(1): 19-22, 2011.

[12] Csermely P, Palotai R and Nussinov R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem Sci*, 35(10): 539-546, 2010.

[13] Frauenfelder H., Sligar S.G. and P.G Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598-603, 1991.

[14] M. Brylinski and J. Skolnick. What is the relationship between the global structures of apo and holo proteins? *Proteins*, 70(2):363-77, 2008.

[15] D. A. Antunes, D. Devaurs and L.E. Kavraki. Understanding the challenges of protein flexibility in drug design. *Expert Opin Drug Discov*, 10(12):1301–1313, 2015.

[16] Najmanovich R., J. Kutter, V. Sobolev and M. Edelman. Side-chain flexibility in proteins upon ligand binding. *Prot. Struct. Funct. Gen.*, 39(3): 261-268, 2000.

[17] S. C. Flores and M. B. Gerstein. Predicting protein ligand binding motions with the conformation explorer. *BMC Bioinformatics*, 12(1):417, 2011.

[18] G. Finocchiaro, T. Wang, R. Hoffmann, A. Gonzalez, and R. C. Wade. Dsmm: a database of simulated molecular motions. *Nucleic Acids Res.*, 31(1): 456-457, 2003.

[19] Babine Robert E. and Bender Steven L. Molecular recognition of protein-ligand complexes: Applications to drug design. *Chemical Reviews (Washington, D. C.)*, 97(5): 1359-1472, 1997.

[20] Z. Huang, L. Zhu, Y. Cao, G. Wu, X. Liu, Y. Chen, Q. Wang, T. Shi, Y. Zhao, Y. Wang, W. Li, Y. Li, H. Chen, G. Chen, and J. Zhang. Asd: a comprehensive database of allosteric proteins and modulators. *Nucleic Acids Res*, 39(Database issue): D663-D669, 2011.

[21] Q. Cui and M. Karplus. Allostery and cooperativity revisited. *Protein Science - A Publication of the Protein Society*, 17(8):1295-307, 2008.

[22] Andrea Basciu, Giuliano Malloci, Fabio Pietrucci, Alexandre M. J. J. Bonvin and Attilio V. Vargiu. Holo-like and druggable protein conformations from enhanced sampling of binding pocket volume and shape. *J. Chem. Inf. Model.*, 59(4):1515–1528, 2019.

[23] D. Seeliger and B. L. De Groot. Conformational transitions upon ligand binding: holo-structure prediction from apo conformations. *PLoS Comput Biol*, 6(1):e1000634, 2010.

[24] Laura Motta, Stefano; Bonati. Modeling binding with large conformational changes: Key points in ensemble-docking approaches. *J Chem Inf Model*, 57(7): 1563-1578, 2017.

[25] Zeynep Kurkcuoglu and Pemra Doruker. Ligand docking to intermediate and close-to-bound conformers generated by an elastic network model based algorithm for highly flexible proteins. *PLOS ONE*, 11(6): e0158063, 2016.

[26] Amaral M., Kokh D. B., Bomke J., Wegener A. et al. Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. *Nat. Commun*, 8(1):2276, 2017.

[27] Hayward S. Identification of specific interactions that drive ligand-induced closure in five enzymes with classic domain movements. *J. Mol. Biol*, 339(4): 1001-1021, 2004.

[28] Wu P., Nielsen T. E. and Clausen M. H. Fda-approved small-molecule kinase inhibitors. *Trends Pharmacol. Sci*, 36(7):422-439, 2015.

[29] Arrowsmith C. H., Bountra C., Fish P. V., Lee K. and Schapira M. Epigenetic protein families: A new frontier for drug discovery. *Nat. Rev. Drug Discovery*, 11(5):384-400, 2012.

[30] Flavin R., Peluso S., Nguyen P. L. and M Loda. Fatty acid synthase as a potential therapeutic target in cancer. *Future Oncol.*, 6(4):551-62, 2010.

[31] Takayuki Amemiya, Ryotaro Koike, Akinori Kidera and Motonori Ota. Pscdb: a database for protein structural change upon ligand binding. *Nucleic Acids Res.*, 40(Database issue): D554–D558, 2011.

[32] Alexander Miguel Monzon, Cristian Oscar Rohr, Maria Silvina Fornasari and Gustavo Parisi. Codnas 2.0: a comprehensive database of protein conformational diversity in the native state. *Database (Oxford)*, 28(2):1-8, 2016.

[33] John Oliver Nealon, Limcy Seby Philomina and Liam James McGuffin. Predictive and experimental approaches for elucidating protein–protein interactions and quaternary structures. *Int. J. Mol. Sci*, 18(12):2623, 2017.

[34] Carroni M and Saibil HR. Cryo electron microscopy to determine the structure of macromolecular complexes. *Methods*, 95(3):78-85, 2016.

[35] Voula Kanelis, Julie D. Forman Kay and Lewis E. Kay. Multidimensional nmr methods for protein structure determination. *IUBMB Life*, 52(6):291-302, 2001.

[36] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235-242, 2000.

[37] A. M. Ruvinsky, T. Kirys, A. V. Tuzikov and I. A. Vakser. Structure flucuations and conformational changes in protein binding. *J Bioinform Comput Biol*, 10(2):1241002, 2012.

[38] Y. Zhang. Progress and challenges in protein structure prediction. *Curr Opin Struct Biol.*, 18(3):342-8, 2009.

[39] Yang Zhang and Jeffrey Skolnick. The protein structure prediction problem could be solved using the current pdb library. *Proc. Natl. Acad. Sci. USA*, 102(4):1029–1034, 2005.

[40] Snyder DA, Chen Y, Denissova NG, Acton T et al. Comparisons of nmr spectral quality and success in crystallization demonstrate that nmr and x-ray crystallography are complementary methods for small protein structure determination. *J Am Chem Soc.*, 127(47):16505-11, 2005.

[41] Joshi A, Esteve V, Buckroyd AN, Blatter M, Allain FH and Curry S. Solution and crystal structures of a c-terminal fragment of the neuronal isoform of the polypyrimidine tract binding protein (nptb). *PeerJ*, 2(1):e305, 2014.

[42] Fernandez C and Wider G. Trosy in nmr studies of the structure and function of large biological macromolecules. *Curr Opin Struct Biol*, 13(5):570-80, 2003.

[43] Wang HW Wang JW. How cryo-electron microscopy and x-ray crystallography complement each other. *Protein Sci.*, 26(1):32-39, 2017.

[44] Rankin NJ, Preiss D, Welsh P, Burgess KE, Nelson SM, Lawlor DA and Sattar N. The emergence of proton nuclear magnetic resonance metabolomics in the cardiovascular arena as viewed from a clinical perspective. *Atherosclerosis*, 237(1):287-300, 2014.

[45] M. Fischer, R. G. Coleman, J. S. Fraser, and B. K. Shoichet. The incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery. *Nat Chem.*, 6(7):575-583, 2014.

[46] Feixas F., Lindert W, Sinko W. and McCammon J.A. Exploring the role of receptor flexibility in structure-based drug discovery. *J Biophys Chem*, 186(4):31-45, 2014.

[47] Lerner E., Cordes T., Ingargiola A., Alhadid Y., Chung S., Michalet X., and Weiss S. Toward dynamic structural biology: Two decades of single-molecule forster resonance energy transfer. *Science*, 359(6373):eaan1133, 2018.

[48] Ward A.B., Sali A. and Wilson I.A. Biochemistry. integrative structural biology. *Science*, 339(6122):913-5, 2013.

[49] S. Mandal, M. Moudgil, and S. K. Mandal. Rational drug design. *Euro J of Pharm*, 625(1-3):90-100, 2009.

[50] M., Hubbard R.E. Renaud J.-P., Chung C., Danielson U.H., Egner U., Hennig and Nar H. Biophysics in drug discovery: impact, challenges and opportunities. *Nat Rev Drug Discov*, 15(10):679-98, 2016.

[51] Zheng H., Hou J., Zimmerman M.D., Wlodawer A. and Minor W. The future of crystallography in drug discovery. *Expert Opin Drug Discov*, 9(2):125-37, 2014.

[52] G. Sliwoski, S. Kothiwale, J. Meiler, and E. W. Lowe. Computational methods in drug discovery. *Pharmacol Rev January*, 66(1): 334–395, 2013.

[53] N. Brooijmans and I. D. Kuntz. Molecular recognition and docking algorithms. *Annu Rev Biophys Biomol Struct*, 32(4):335-73, 2003.

[54] X.-Y. Meng, H.-X. Zhang, M. Mezei, and M. Cui. Molecular docking: A powerful approach for structure-based drug discovery. *Curr Comput Aided Drug Des.*, 7(2): 146-157, 2011.

[55] M. Karplus and A. McCammon. Molecular dynamics simulations of biomolecules. *Nat Struct Bio*, 9(3): 646-788, 2002.

[56] Adam Hospital, Josep Ramon Goni, Modesto Orozco and Josep L Gelpi. Molecular dynamics simulations: advances and applications. *Adv. Appl. Bioinform. Chem.*, 8(3): 37-47, 2015.

[57] C. Best and H. Hege. Visualizing and identifying conformational ensembles in molecular dynamics trajectories. *Comp in Sci Eng*, 4(3):68–75, May 2002.

[58] Hart TN and Read RJ. A multiple-start monte carlo docking method. *Proteins*, 13(3):206-22, 1992.

[59] Maurizio R. Buonfiglio, M. Recanatini and M. Masetti. Protein flexibility in drug discovery: From theory to computation. *ChemMedChem*, 10(7):1141–1148, 2015.

[60] M. De Vivo, M. Masetti, G. Bottegoni, A. Cavalli. Role of molecular dynamics and related methods in drug discovery. *J Med Chem*, 59(9):4035–4061, 2016.

[61] Uehara Shota and Shigenori Tanaka. Cosolvent-based molecular dynamics for ensemble docking: Practical method for generating druggable protein conformations. *J. Chem. Inf. Model*, 57(4):742-756, 2017.

[62] Hamelberg Donald, Mongan John, and McCammon J. Andrew. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J Chem Phys*, 120(24):11919-29, 2004.

[63] M. Masetti, A. Cavalli, M. Recanatini, and F. L. Gervasio. Exploring complex protein-ligand recognition mechanisms with coarse metadynamics. *J Phys Chem B*, 113(14):4807-16, 2009.

[64] G. Bussi, A. Laio, and M. Parrinello. Equilibrium free energies from non-equilibrium metadynamics. *Phys. Rev. Lett.*, 96: 090601, 2006.

[65] Miao Yinglong, Goldfeld Dahlia Anne, and Von Moo Ee et al. Accelerated structure-based design of chemically diverse allosteric modulators of a muscarinic g protein-coupled receptor. *Proc Natl Acad Sci U S A*, 113(38): E5675-E5684, 2016.

[66] C. Abrams and G. Bussi. Enhanced sampling in molecular dynamics using metadynamics, replica exchange, and temperature acceleration. *Entropy*, 16(1): 163-199, 2014.

[67] V. Mohan, A. C. Gibbs, M. D. Cummings, E. P. Jaeger, and R. L. DesJarlais. Docking: Successes and challenges. *Curr Pharm Des*, 11(3): 323-33, 2005.

[68] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, 3(11):935-49, 2004.

[69] Y. C. Chen. Beware of docking! *Trends Pharmacol Sci.*, 36(2):78-95, 2015.

[70] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: An overviewof search algorithms and a guide to scoring functions. *PROTEINS: Structure, Function, and Genetics*, 47(4):409-43, 2002.

[71] Irwin J.J. and Shoichet B.K. Docking screens for novel ligands conferring new biology. *J Med Chem*, 59(9):4103-20, 2016.

[72] L. G. Ferreira, R. N. dos Santos, G. Oliva, A. D. Andricopulo. Molecular docking and structure-based drug design strategies. *Molecules*, 20(7):13384–13421, 2015.

[73] Forli S. Charting a path to success in virtual screening. *Molecules*, 20(10):18732-58, 2015.

[74] Ronnie Amaro, J. Baudry, J. Chodera, O. Demir, J. A. McCammon, Y. Miao, J. C. Smith. Ensemble docking in drug discovery. *Biophys J*, 114(10):2271–2278, 2018.

[75] Nataraj S. Pagadala, Khajamohiddin Syed and Jack Tuszynski. Software for molecular docking: a review. *Biophys Rev*, 9(2):91-102, 2017.

[76] Ferrari AM, Wei BQ, Costantino L, Shoichet BK. Soft docking and multiple receptor conformations in virtual screening. *J Med Chem*, 47(21): 5076-5084, 2004.

[77] Jiang F and Kim SH. Soft docking: Matching of molecular surface cubes. *J Mol Biol*, 219(1):79-102, 1991.

[78] Xiaoqin Huang, Sheng-You; Zou. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins*, 66(2):399–421, 2007.

[79] O. Korb, T. S. G. Olsson, S. J. Bowden, R. J. Hall, M. L. Verdonk, J. W. Liebeschuetz, and J. C. Cole. Potential and limitations of ensemble docking. *J. Chem. Inf. Model.*, 52(5): 1262-1274, 2012.

[80] A. Laio and M. Parrinello. Escaping free-energy minima. *PNAS*, 99(20):12562-66, 2002.

[81] Jie Ping, Pei Hao, Yi-Xue Li, and Jing-Fang Wang. Molecular dynamics studies on the conformational transitions of adenylate kinase: A computational evidence for the conformational selection mechanism. *Biomed res int*, 2013: 628536, 2013.

[82] E. Formoso, V. Limongelli, and M. Parrinello. Energetics and structural characterization of the large-scale functional motion of adenylate kinase. *Sci Rep*, 5:8425, 2015.

[83] Dechang Li, Ming S.Liu and Baohua J. Mapping the dynamics landscape of conformational transitions in enzyme: The adenylate kinase case. *Biophys J*, 109(3):647-60, 2015.

[84] P. Atkins, J. de Paula, and R. Friedman. *Quanta, Matter and Change: A Molecular Approach to Physical Chemistry.* W. H. Freeman; First Edition edition, 2008.

[85] P. L. Kastritis and A. M. J. J. Bonvin. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *J. R. Soc. Interface*, 10(79):20120835, 2013.

[86] M K Gilson, J A Given, B L Bush and J A McCammon. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys J*, 72(3): 1047–1069, 1997.

[87] Ronald J. Greaves and Kenneth D. Schlecht. Gibbs free energy: The criteria for spontaneity. *J. Chem. Educ*, 69(5): 417, 1992.

[88] Perozzo R, Folkers G and Scapozza L. Thermodynamics of protein-ligand interactions: history, presence, and future aspects. *J Recept Signal Transduct Res*, 24(1-2):1-52, 2004.

[89] Maria Luisa Verteramo, Olof Stenstrom, Majda Misini Ignjatovic, Octav Caldararu, Martin A. Olsson, et al. Interplay between conformational entropy and solvation entropy in protein–ligand binding. *J. Am. Chem. Soc*, 141(5): 2012-2026, 2019.

[90] Di Cui, Bin W. Zhang, Orcid Nobuyuki Matubayasi, Ronald M. Levy. The role of interfacial water in protein–ligand binding: Insights from the indirect solvent mediated potential of mean force. *J. Chem. Theory Comput.*, 14(2): 512-526, 2018.

[91] Michelle A. Sahai and Philip C. Biggin. Quantifying water-mediated protein–ligand interactions in a glutamate receptor: A dft study. *J Phys Chem B*, 115(21): 7085-7096, 2011.

[92] J. Michel, J. Tirado-Rives, and W. L. Jorgensen. Energetics of displacing water molecules from protein binding sites: consequences for ligand optimization. *J. Am. Chem. Soc.*, 131(42): 15403-15411, 2009.

[93] Deliang Chen, Numan Oezguen, Petri Urvil, Colin Ferguson, Sara M. Dann, and Tor C. Savidge. Regulation of protein-ligand binding affinity by hydrogen bond pairing. *Sci Adv*, 2(3):e1501240, 2016.

[94] Chodera JD and Mobley DL. Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Annu Rev Biophys*, 42(1):121-42, 2013.

[95] Barillari C, Taylor J, Viner R and Essex JW. Classification of water molecules in protein binding sites. *J Am Chem Soc*, 129(9): 992577-2587, 2007.

[96] Mac Raild C. A., Daranas A. H. , Bronowska A. and Homans S. W. Global changes in local protein dynamics reduce the entropic cost of carbohydrate binding in the arabinose-binding protein. *J Mol Biol*, 368(3):822-32, 2007.

[97] Ulf Ryde. A fundamental view of enthalpy–entropy compensation. *Med. Chem. Commun.*, 5(3): 1324-1336, 2014.

[98] Piotr Setny, Riccardo Baron and J. Andrew McCammon. How can hydrophobic association be enthalpy driven? *J Chem Theory Comput*, 6(9): 2866-2871, 2010.

[99] Johannes Schiebel, Roberto Gaspari, Tobias Wulsdorf, Khang Ngo et al. Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes. *Nat Commun*, 9(1):3559, 2018.

[100] Song-Ho Chong and Sihyun Ham. Dynamics of hydration water plays a key role in determining the binding thermodynamics of protein complexes. *Sci Rep*, 7(1): 8744, 2017.

[101] John E Ladbury. Just add water! the effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem Biol*, 3(12): 973-980, 1996.

[102] Fisher E. Einfluss der configuration auf die wirkung der enzyme. *Ber. Dtsch. Chem. Ges*, 27(3):2985-2993, 1984.

[103] Amit AG, Mariuzza RA, Phillips SE and Poljak RJ. Three-dimensional structure of an antigen-antibody complex at 2.8 a resolution. *Science*, 233(4765):747-53, 1986.

[104] David M. Blow. Structure and mechanism of chymotrypsin. *Acc. Chem. Res*, 9(4): 145-152, 1976.

[105] D. E. Koshland Jr. Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci U S A.*, 44(2): 98–104, 1958.

[106] Wlodawer A and Vondrasek J. Inhibitors of hiv-1 protease: a major success of structure-assisted drug design. *Annu Rev Bioph Biom*, 27:249-84, 1998.

[107] Bystroff C and Kraut J. Crystal structure of unliganded escherichia coli dihydrofolate reductase. ligand-induced conformational changes and cooperativity in binding. *Biochemistry-US*, 30(8): 2227-2239, 1991.

[108] Zhao S, Goodsell DS and Olson AJ. Analysis of a data set of paired uncomplexed protein structures: new metrics for side-chain flexibility and model evaluation. *Proteins*, 43(3):271-9, 2001.

[109] Vazquez-Laslop N, Zheleznova EE, Markham PN, Brennan RG and Neyfakh AA. Recognition of multiple drugs by a single protein: a trivial solution of an old paradox. *Biochem Soc T*, 43(3):271-9, 2000.

[110] Teodoro and L. E. Kavraki. Conformational flexibility models for the receptor in structure based drug design. *Curr Pharm Des*, 9(20):1635-48, 2003.

[111] C.J. Ma, B., Kumar S., Tsai and Nussinov R. Folding funnels and binding mechanisms. *Protein Eng*, 12(9):713-720, 1999.

[112] H.-J. Böhm and G. Schneider. *Protein-Ligand Interactions: From Molecular Recognition to Drug Design*, volume 19. Wiley-VCH, 2006.

[113] Straub FB and Szabolcsi G. Remarks on the dynamic aspect of enzyme structure. *Molecular biology: problems and perspectives*, 272(915):109-122, 1964.

[114] Tsai CD, Ma B, Kumar S, Wolfson H and Nussinov R. Protein folding: binding of conformationally fluctuating building blocks via population selection. *Crit Rev Biochem Mol Biol*, 36(5):399-433, 2001.

[115] S. Kumar, B. Ma, C. J. Tsai, N. Sinha, and R. Nussinov. Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci.*, 9(1):10-9, 2000.

[116] Henzler-Wildman K.A. and Kern D. Dynamic personalities of proteins. *Nature*, 450(7172): 964–972, 2007.

[117] Bryngelson J.D., Onuchic J.N., Socci N.D. and Wolynes P.G. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins*, 21(3):167-95, 1995.

[118] Miller D.W. and Dill K.A. Ligand binding to proteins: The binding landscape model. *Protein Sci*, 6(10): 2166–2179, 1997.

[119] X. Salvatella. Understanding protein dynamics using conformational ensembles. *Adv Exp Med Biol.*, 805(9):67-85.

[120] V. N. Uversky. Under-folded proteins: Conformational ensembles and their roles in protein folding, function, and pathogenesis. *Biopolymers.*, 99(11):870-87, 2013.

[121] J. R. Kasper and C. Park. Ligand binding to a high-energy partially unfolded protein. *Protein Sci.*, 24(1):129-37, 2015.

[122] Weikl TR and Paul F. Conformational selection in protein binding and function. *Protein Sci*, 23(11):1508-18, 2014.

[123] Hammes GG1, Chang YC and Oas TG. Conformational selection or induced fit: a flux description of reaction mechanism. *Proc Natl Acad Sci U S A*, 106(33):13737-41, 2009.

[124] Grunberg R., Leckner J. and Nilges M. Complementarity of structure ensembles in protein-protein binding. *Structure*, 12(12):2125-36, 2004.

[125] Wlodarski T. and Zagrovic B. Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. *Proc. Natl. Acad. Sci*, 106(46):19346-19351, 2009.

[126] M. Iskar and G. Zeller et al. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. *Curr Opin Biotechnol*, 23(4):609-16, 2011.

[127] R. Petrenko and J. Meller. Molecular dynamics. *eLS. John Wiley and Sons Ltd, Chichester*, 2010.

[128] M. E. Tuckerman and G. J. Martyna. Understanding modern molecular dynamics: Techniques and applications. *J. Phys. Chem. B*, 104(2):159–178, 2000.

[129] D. L. Beveridge and F. M. DiCapua. Free energy via molecular simulation: applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Chem.*, 18(4):431-92, 1989.

[130] C. Lubich. *From Quantum to Classical Molecular Dynamics: Reduced Models and Numerical Analysis.* Zurich Lectures in Advanced Mathematics. European Mathematical Society, 2008.

[131] D. Frenkel and B. Smit. *Understanding Molecular Simulation, Second Edition: From Algorithms to Applications.* Academic Press; 2 edition, 2001.

[132] M. Griebel, S. Knapek, and G. Zumbusch. *Numerical Simulation in Molecular Dynamics: Numerics, Algorithms, Parallelization, Applications.* Springer Publishing Company, 2007.

[133] Edmond Chow, Xing Liu, Mikhail Smelyanskiy, Jeff R. Hammond. Parallel scalability of hartree–fock calculations. *J. Chem. Phys*, 142(10):104103, 2015.

[134] Marilia T. C. Martins Costa, Manuel F. Ruiz Lopez. Reaching multi nanosecond timescales in combined qm/mm molecular dynamics simulations through parallel horsetail sampling. *J Comput Chem*, 38(10):659-668, 2017.

[135] O. M. Becker, A. D. MacKerell, B. Roux, and M. Watanabe. *Computational Biochemistry and Biophysics.* CRC Press; 1st edition, 2001.

[136] D. A. Case, T. E. Cheathman, T. Darden, H. Gohlke, R. Luo, K. M. Merz, and R. J. Woods. The amber biomolecular simulation programs. *J Comput Chem*, 26(16):1668-88, 2005.

[137] D. Boda, D. Henderson. The effects of deviations from lorentz–berthelot rules on the properties of a simple mixture. *Molecular Physics*, 106(20):2367-2370, 2008.

[138] P. Dauber-Osguthorpe, T. Hagler. Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there? *J Comput Aided Mol Des*, 33(2):133-203, 2019.

[139] Zhao H, Caflisch A. Molecular dynamics in drug design. *Eur J Med Chem*, 16(91):4-14, 2015.

[140] A. Leach. *Molecular Modelling: Principles and Applications.* Prentice Hall; 2 edition, 2001.

[141] L. Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev.,*, 159(1): 98–103, 1967.

[142] H. Grubmüller and P. Tavan. Multiple time step algorithms for molecular dynamics simulations of proteins: How good are they? *J Comput Chem*, 19(13): 1534-1552, 1998.

[143] C. Jong-In and K. Byungchul. Determination of proper time step for molecular dynamics simulation. *Bull. Korean Chem. Soc.*, 21(4):419-424, 2000.

[144] Walker Ross C. Hopkins Chad W, Le Grand Scott and Roitberg Adrian E. Long-time-step molecular dynamics through hydrogen mass repartitioning. *J Chem Theory Comput*, 11(4):1864-1874, 2015.

[145] W. C. Swope, H.C. Andersen, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J Chem Phys*, 76(1):637-649, 1982.

[146] R. Hockney, S. Goel, and J. Eastwood. Quiet high-resolution computer models of a plasma. *J Comp Phys*, 14(2):148-158, 1974.

[147] C. Chipot and A. Pohorille, editors. *Free Energy Calculations*. Springer-Verlag Berlin Heidelberg, 2007.

[148] Jaewoon Jung, Wataru Nishima, Marcus Daniels et al. Scaling molecular dynamics beyond 100,000 processor cores for large scale biophysical simulations. *J Comput Chem*, 40(21):1919-1930, 2019.

[149] P. P. Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. *Ann. Phys.*, 369(3):253-287, 1921.

[150] D. York T. Darden and L. Pedersen. Particle mesh ewald: An nlog(n) method for ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10093, 1993.

[151] P. A. Bash, U. C. Singh, F. K. Brown, R. Langridge, and P. A. Kollman. Free energy calculation by computer simulation. *Science*, 236(4801):574–576, 1987.

[152] B. Ensing, M. De Vivo, Z. Liu, P. Moore, and M. L. Klein. Metadynamics as a tool for exploring free energy landscapes of chemical reactions. *Acc Chem Res.*, 39(2):73-81, 2006.

[153] B. Ensing, A. Laio, M. Parrinello, and M. L. Klein. A recipe for the computation of the free energy barrier and the lowest free energy path of concerted reactions. *J Phys Chem B.*, 109(14):6676-87, 2005.

[154] A. Barducci, M. Bonomi, and M. Parrinello. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5): 826-843, 2011.

[155] O. J. Andersen, J. Grouleff, P. Needham, R. C. Walker, and F. Jensen. Toward an enhanced sampling molecular dynamics method for studying ligand-induced conformational changes in proteins. *J. Phys. Chem. B*, 119(46):14594-14603, 2015.

[156] Rafael C.Bernardia , Marcelo C.R.Melo and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochem Biophys acta*, 1850(5):872-877, 2015.

[157] Chou KC, Carlacci L. Simulated annealing approach to the study of protein structures. *Protein Eng*, 4(6):661-7, 1991.

[158] Brunger AT, Adams PD, Rice LM. New applications of simulated annealing in x-ray crystallography and solution nmr. *Structure*, 5(3):325-36, 1997.

[159] Y. Sugita, Y.; Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett*, 314(1,2): 141-151, 1999.

[160] Luitz M, Bomblies R, Ostermeir K, Zacharias M. Exploring biomolecular dynamics and interactions using advanced sampling methods. *J Phys Condens Matter*, 27(32):323101, 2015.

[161] David J. Earl and Michael W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Phys Chem Chem Phys*, 7(23):3910-3916, 2005.

[162] M. Meli and G. Colombo. A hamiltonian replica exchange molecular dynamics (md) method for the study of folding, based on the analysis of the stabilization determinants of proteins. *Int J Mol Sci*, 14(6): 12157–12169, 2013.

[163] Katja Ostermeir and Martin Zacharias. Advanced replica-exchange sampling to study the flexibility and plasticity of peptides and proteins. *Biochimica et Biophysica Acta*, 1834(5):847-853, 2013.

[164] Martin Luitz, Manuel P.; Zacharias. Protein-ligand docking using hamiltonian replica exchange simulations with soft core potentials. *J Chem Inf Model*, 54(6):1669–1675, 2014.

[165] F. Mandl. *Statistical Physics*. Wiley; 2 edition, 1988.

[166] S. Kumar, I. D. Bouzida, R. H. Swendsen, P. A. Kollman PA, and J. M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules i. the method. *J. Comput. Chem.*, 13(8):1011-1021, 1992.

[167] A. Pandini, A. Fornili. Using local states to drive the sampling of global conformations in proteins. *J. Chem. Theory Comput*, 12(3):1368-79, 2016.

[168] P. Raiteri, A. Laio, F. L. Gervasio, C. Micheletti, and M. Parrinello. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J. Phys. Chem. B*, 110(8): 3533-3539, 2006.

[169] F. Marinelli, F. Pietrucci, A. Laio, and S. Piana. A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. *PLoS Comput Biol*, 5(8):e1000452, 2009.

[170] Laio A., Rodriguez Fortea A., Gervasio F. L., Ceccarelli M. and M. Parrinello. Assessing the accuracy of metadynamics. *J. Phys. Chem. B*, 109(14):6714-21, 2005.

[171] A. Barducci, G. Bussi, and M. Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Physical review letters*, 100(2): 020603, 2008.

[172] S. Piana and A. Laio. A bias-exchange approach to protein folding. *J. Phys. Chem. B*, 111(17):4553-9, 2007.

[173] Iglesias J., Saen oon S., Soliva R. and V. Guallar. Computational structure-based drug design: Predicting target flexibility. *WIREs Comput Mol Sci*, 8(5):e1367, 2018.

[174] C. Beier and M. Zacharias. Tackling the challenges posed by target flexibility in drug design. *Opinion on Drug Discovery*, 5(4):347-59, 2010.

[175] V. Limongelli, L. Marinelli, S. Cosconati et al. Sampling protein motion and solvent effect during ligand binding. *PNAS*, 109(5):1467-1472, 2012.

[176] Shoichet BK, Leach AR and Kuntz ID. Ligand solvation in molecular docking. *Proteins*, 34(1):4-16, 1999.

[177] Sergey A. Samsonov, Joan Teyra and M. Teresa Pisabarro. Docking glycosaminoglycans to proteins: analysis of solvent inclusion. *J Comput Aided Mol Des*, 25(5):477-89, 2011.

[178] J. Apostolakis, A. Pluckthun and Amedeo Caflisch. Docking small ligands in flexible binding sites. *J Comput Chem*, 19(1):21-37, 1998.

[179] Mizutani MY, Takamatsu Y, Ichinose T et al. Effective handling of induced-fit motion in flexible docking. *Proteins*, 63(4):878-91, 2006.

[180] Sandak B, Wolfson HJ and Nussinov R. Flexible docking allowing induced fit in proteins: Insights from an open to closed conformational isomers. *Proteins*, 32(2):159-74, 1998.

[181] Lina Kulakova, Georgios Arampatzis, Panagiotis Angelikopoulos, Panagiotis Hadjidoukas, et al. Data driven inference for the repulsive exponent of the lennard-jones potential in molecular dynamics simulations. *Sci Rep*, 7(2):16576, 2017.

[182] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, and R. D. Taylor. Improved protein-ligand docking using gold. *Proteins*, 52(4):609-23, 2003.

[183] B-Rao C, Subramanian J and Sharma SD. Managing protein flexibility in docking and its applications. *Drug Discov Today*, 14(7-8):394-400, 2009.

[184] Sherman W, Day T, Jacobson MP et al. Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem*, 49(2):534-53, 2006.

[185] Venkatraman V and Ritchie DW. Flexible protein docking refinement using pose-dependent normal mode analysis. *Proteins*, 80(9):2262-74, 2012.

[186] Desmet J, Wilson IA, Joniau M et al. Computation of the binding of fully flexible peptides to proteins with flexible side chains. *FASEB*, 11(2):164-72, 1997.

[187] J. Meiler and D. Baker. Rosettaligand: protein-small molecule docking with full side-chain flexibility. *PROTEINS: Struct, Funct and Bioinf*, 65: 538–548, 2006.

[188] Flick J, Tristram F and Wenzel W. Modeling loop backbone flexibility in receptor-ligand docking simulations. *J Comput Chem*, 33(31):2504-15, 2012.

[189] Ding F, Yin S and Dokholyan NV. Rapid flexible docking using a stochastic rotamer library of ligands. *J Chem Inf Model*, 50(9):1623-1632, 2010.

[190] Shin WH and Seok C. Galaxydock: Protein-ligand docking with flexible protein side-chains. *J Chem Inf Model*, 52(12):3225-32, 2012.

[191] Schumann M and Armen RS. Systematic and efficient side chain optimization for molecular docking using a cheapest-path procedure. *J Comput Chem*, 34(14):1258-69, 2013.

[192] A. May and M. Zacharias. Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking. *Proteins*, 70(3):794-809, 2007.

[193] Sheng-You Huang and Xiaoqin Zou. Advances and challenges in protein-ligand docking. *Int. J. Mol. Sci*, 11(8):3016-34, 2010.

[194] H. Alonso, A. A. Bliznyuk, and J. E. Gready. Combining docking and molecular dynamic simulations in drug design. *Med Res Rev*, 26(5):531-568, 2006.

[195] Woody Sherman, Hege S. Beard and Ramy Farid. Use of an induced fit receptor structure in virtual screening. *Chem Bio Drug Des*, 67(1):83-4, 2005.

[196] Veronica Salmaso and Stefano Moro. Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. *Front Pharmacol*, 9(5):923, 2018.

[197] Jacobson M. P., Friesner R.A., Xiang Z. and Honig B. On the role of crystal packing forces in determining protein sidechain conformations. *J. Mol. Biol*, 320(3):597-608, 2002.

[198] P. Jacobson, David L. Pincus, Chaya S. Rapp, Tyler J.F. Day, Barry Honig et al. A hierarchical approach to all atom protein loop prediction. *Proteins*, 55(2):351-67, 2004.

[199] Borrelli KW, Cossins B and Guallar V. Exploring hierarchical refinement techniques for induced fit docking with protein and ligand flexibility. *J Comput Chem*, 31(6):1224-35, 2010.

[200] C Hartmann, R. Banisch, M. Sarich, T. Badowski, and C. Schutte. Characterization of rare events in molecular dynamics. *Entropy*, 16(1):350-376, 2014.

[201] Ronnie Amaro, W. Wilfred Li. Emerging methods for ensemble-based virtual screening. *Current Topics in Medicinal Chemistry*, 10(1):3-13, 2010.

[202] Kurkcuoglu Zeynep, Koukos Panagiotis I., Citro Nevia, and Trellet Mikael E. et al. Performance of haddock and a simple contact-based protein-ligand binding affinity predictor in the d3r grand challenge 2. *J Comput Aided Mol Des*, 32(1):175-185, 2018.

[203] A. Ganesan, M. Coote, K. Barakat. Molecular dynamics-driven drug discovery: leaping forward with confidence. *Drug Discovery Today*, 22(2):249–269, 2017.

[204] Colin R. Hopkins, Andrew L.; Groom. Opinion: The druggable genome. *Nature Reviews Drug Discovery*, 1(9):727–730, 2002.

[205] Chris Finan, Anna Gaulton, Felix. A Kruger, R. Thomas Lumbers, Tina Shah, Jorgen Engmann, Luana Galver, Ryan Kelley, Anneli Karlsson, Rita Santos, John P. Overington, Aroon D. Hingorani and Juan P. Casas. The druggable genome and support for target identification and validation in drug development. *Sci Transl Med.*, 9(383):1166, 2017.

[206] C. N. Cavasotto, J. A. Kovacs, and R. A. Abagyan. Representing receptor flexibility in ligand docking through relevant normal modes. *J Am Chem Soc*, 127(26): 9632–9640, 2005.

[207] R. T. Kroemer. Structure-based drug design: Docking and scoring. *Current Protein and Peptide Science*, 8(1): 312-328, 2007.

[208] Lin Jung-Hsin, Perryman Alexander L, Schames Julie R, and McCammon J Andrew. Computational drug design accommodating receptor flexibility: The relaxed complex scheme. *J Am Chem Soc*, 124(20):5632–5633, 2002.

[209] A. Tarcsay, G. Paragi, M. Vas, B. Jojart, F. Bogar, G. M. Keseru. The impact of molecular dynamics sampling on the performance of virtual screening against gpcrs. *J Chem Inf Model*, 53(11):2990-2999, 2013.

[210] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, Jun 2015.

[211] Rao Huang, Li-Ta Lo, Yuhua Wen, Arthur F. Voter, and Danny Perez. Cluster analysis of accelerated molecular dynamics simulations: A case study of the decahedron to icosahedron transition in pt nanoparticles. *J Chem Phys*, 147(15):152717, 2017.

[212] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

[213] Jianyin Shao, Stephen W. Tanner, Nephi Thompson, and Thomas E. Cheatham. Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms. *J Am Chem Soc*, 3(6):2312–2334, 2007.

[214] P.Baby K.Sasirekha. Agglomerative hierarchical clustering algorithm- a review. *Int J Sci Res Pub*, 3(3):86-97, 2013.

[215] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

[216] Steinhaus H. Sur la division des corps matériels en parties. *Bulletin de l'Académie Polonaise des Sciences. 1956*, 4(12):801–804, 1956.

[217] Joaquín Pérez-Ortega, Nelva Nely Almanza-Ortega, and David Romero. Balancing effort and benefit of k-means clustering algorithms in big data realms. *PLOS ONE*, 13(9):1–19, 2018.

[218] Lloyd S. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[219] Jancey RC. Multidimensional group analysis. *Austr J Bot*, 14(1):127–130, 1966.

[220] Osguthorpe David J, Sherman Woody and Hagler Arnold T. Generation of receptor structural ensembles for virtual screening using binding site shape analysis and clustering. *Chem Bio Drug Des*, 80(2):182–193, 2012.

[221] T. Maximova, R. Moffatt, B. Ma, R. Nussinov and A. Shehu. Principles and overview of sampling methods for modeling macromolecular structure and dynamics. *PLoS Comput Biol*, 12(4): e1004619, 2016.

[222] Huang SY, Grinter SZ and Zou X. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Phys Chem Chem Phys*, 12(40):12899-908, 2010.

[223] Jain AN. Scoring functions for protein-ligand docking. *Curr Protein Pept Sci*, 7(5):407-20, 2006.

[224] William L. Jorgensen and Laura L. Thomas. Perspective on free-energy perturbation calculations for chemical equilibria. *J Chem Theory Comput*, 4(6):869-876, 2008.

[225] Robert W. Zwanzig. High temperature equation of state by a perturbation method. i. nonpolar gases. *J. Chem. Phys*, 22(1):1420, 1954.

[226] Straatsma TP and McCammon JA. Theoretical calculations of relative affinities of binding. *Methods Enzymol*, 202(1):497-511, 1991.

[227] S. Genheden and U. Ryde. The mm/pbsa and mm/gbsa methods to estimate ligand-binding affinities. *Expert Opin Drug Discov.*, 10(5):449-61, 2015.

[228] El Khoury Lea, Santos-Martins Diogo, Sasmal Sukanya, Eberhardt Jerome, Bianco Giulia, Ambrosio Francesca Alessandra, et al. Comparison of ligand affinity ranking using autodock-gpu and mm-gbsa scores in the d3r grand challenge 4. *ChemRxiv. Preprint*, 2019.

[229] Liu J, Wang R. Classification of current scoring functions. *J Chem Inf Model*, 55(3):475-82, 2015.

[230] Huang S.Y. and Zou X. Mean-force scoring functions for protein-ligand binding. *Annu. Rep. Comput. Chem*, 6(1):280-296, 2010.

[231] Verma J, Khedkar VM and Coutinho EC. 3d-qsar in drug design–a review. *Curr Top Med Chem*, 10(1):95-115, 2010.

[232] Santiago Vilar and Stefano Costanzi. Predicting biological activities through qsar analysis and docking-based scoring. *Methods Mol Biol*, 914(1):271-84, 2012.

[233] Markus A. Lill. Multi-dimensional qsar in drug discovery. *Drug Discovery Today*, 12(23-24):1013-7, 2007.

[234] Christiane Ehrt, Tobias Brinkjost and Oliver Koch. Binding site characterization - similarity, promiscuity, and druggability. *Med. Chem. Commun.*, 10(1):1145-1159, 2019.

[235] Xavier Barril. Druggability predictions: methods, limitations, and applications. *WIREs*, 3(4):327-338, 2012.

[236] V. Le Guilloux, P. Schmidtke, and P. Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, 10(1):168, 2009.

[237] Schmidtke Peter and Barril Xavier. Understanding and predicting druggability. a high-throughput method for detection of drug binding sites. *J Med Chem*, 53(15):5858-5867, 2010.

[238] Irina Kufareva and Ruben Abagyan. Methods of protein structure comparison. *Methods Mol Biol*, 875(1):231-257, 2012.

[239] Vytautas Gapsys and Bert L. de Groot. Optimal superpositioning of flexible molecule ensembles. *Biophys J*, 104(1):196–207, 2013.

[240] Kabsch W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, 32(1):922–923, 1976.

[241] Kabsch W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, 34(1):827–828, 1978.

[242] Horn B.K.P. Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Am. A Opt. Image Sci. Vis*, 4(1):629–642, 1987.

[243] Mary E. Karpen, Douglas J. Tobias, and Charles L. Brooks. Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of ypgdv. *Biochem*, 32(2):412–420, 01 1993.

[244] Andre G. Michel and Catherine Jeandenans. Multiconformational investigations of polypeptidic structures, using clustering methods and principal components analysis. *Comput and Chem*, 17(1):49 – 59, 1993.

[245] Marja T. Hyvönen, Yrjö Hiltunen, Wael El-Deredy, Timo Ojala, Juha Vaara, Petri T. Kovanen, and Mika Ala-Korpela. Application of self-organizing maps in conformational analysis of lipids. *J Am Chem Soc*, 123(5):810–816, 02 2001.

[246] Daniel R. Roe and Thomas E. Cheatham. Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput*, 9(7):3084–3095, 2013.

[247] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. Gromacs: A message-passing parallel molecular dynamics implementation. *Comp Phys Comm*, 91(13):43 – 56, 1995.

[248] S.C.Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(1):241-254, 1967.

[249] J. H. Ward. Hierarchical grouping to optimize an objective function. *J Am Stat Ass*, 58(1):236–244, 1963.

[250] Fionn Murtagh. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *J Classif*, 31(1):274-295, 2014.

[251] Franz Aurenhammer. Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3):345–405, 1991.

[252] J. A. Hartigan and M. A Wong. Algorithm as 136: A k-means clustering algorithm. *J R Stat Soc*, 28(1):100–108, 1979.

[253] E. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21(1):768–780, 1965.

[254] Greg Hamerly and Charles Elkan. Alternatives to the k-means algorithm that find better clusterings. *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, 8(1):600–607, 2002.

[255] M. Emre Celebi, Hassan Kingravi, and Patricio A. Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1):200–210, 2013.

[256] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[257] Qi Wu, Zhenling Peng, Yang Zhang and Jianyi Yang. Coach-d: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Res.*, 46(W1):W438-W442, 2018.

[258] Schrodinger. Maestro. *Release*, 2019-2.

[259] Morera Solange, Lariviere Laurent, Kurzeck Jurgen, Aschke-Sonnenborn Ursula, Freemont Paul S., Janin Joel, and Ruger Wolfgang. High resolution crystal structures of t4 phage beta-glucosyltransferase: Induced fit and effect of substrate and metal binding. *J Mol Biol*, 311(3):569-77, 2001.

[260] de Groot B. L, van Aalten D. M. F., Scheek R. M., Amadei A., Vriend G., and Berendsen H. J. C. Prediction of protein conformational freedom from distance constraints. *Proteins: Structure, Function, and Genetics*, 29(2):240-51, 1997.

[261] Haas Juergen Seeliger Daniel and de Groot Bert L. Geometry-based sampling of conformational transitions in proteins. *Structure (Cambridge, MA, United States)*, 15(11):1482-92, 2007.

[262] Misna Debra, Monzingo Arthur F., Katzin Betsy J., Ernst Stephen, and Robertus Jon D. Structure of recombinant ricin a chain at 2.3 angstrom. *Protein Science*, 2(3):429-35, 1993.

[263] Yan Xinjian, Hollis Thomas, Svinth Maria, Day Philip, Monzingo Arthur F., Milne George W. A., and Robertus Jon D. Structure-based identification of a ricin inhibitor. *J Mol Biol*, 266(5):1043-9, 1997.

[264] M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. M. Mooij, P. N. Mortenson, and C. W. Murray. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.*, 50(4):726–741, 2007.

[265] Ravindranath Pradeep Anand, Forli Stefano, Goodsell David S., Olson Arthur J., and Sanner Michel F. Autodockfr: advances in protein-ligand docking with explicitly specified binding site flexibility. *PLoS Comput Biol*, 11(12):e1004586, 2015.

[266] O. Trott and A. J. Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J Comput Chem.*, 31(2): 455–461, 2010.

[267] Jones T. Alwyn Chaudhuri Barnali Neel, Ko Junsang; Park Chankyu and Mowbray Sherry L. Structure of d-allose binding protein from escherichia coli bound to d-allose at 1.8 angstrom resolution. *J Mol Biol*, 286(5):1519-31, 1999.

[268] Thomas A. Halgren, Robert B. Murphy, Richard A. Friesner, Hege S. Beard, Leah L. Frye, W. Thomas Pollard, and Jay L. Banks. Glide: A new approach for rapid, accurate docking and scoring. *J. Med. Chem.*, 47(7):1750–1759, 2004.

[269] Friesner Richard A., Murphy Robert B., Repasky Matthew P., Frye Leah L., Greenwood Jeremy R., Halgren Thomas A., Sanschagrin Paul C., and Mainz Daniel T. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem*, 49(21):6177-96, 2006.

[270] C. Dominguez, R. Boelens, and A. M.J. J. Bonvin. Haddock: A protein-protein docking approach based on biochemical or biophysical information. *Biophys J*, 125(7):1731-7, 2003.

[271] Morris Garrett M, Huey Ruth, Lindstrom William, Sanner Michel F, Belew Richard K., Goodsell David S., and Olson Arthur J. Autodock and autodocktools: Automated docking with selective receptor flexibility. *J Comput Chem*, 30(16):2785-91, 2009.

[272] Case D. A., Betz R. M., Cerutti D. S., Cheatham T. E. III, Darden T. A., Duke R. E., Giese T. J., Gohlke H., Goetz A. W., Homeyer N., and Izadi S. et al. Amber16. *University of California: San Francisco*, 2016.

[273] Maier James A., Martinez Carmenza, Kasavajhala Koushik, Wickstrom Lauren, Hauser Kevin E., and Simmerling Carlos. ff14sb: Improving the accuracy

of protein side chain and backbone parameters from ff99sb. *J Chem Theory Comput*, 11(8):3696-713, 2015.

[274] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(1):1157-1174., 2004.

[275] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, and J. R. Cheeseman et al. Gaussian 09, revision a.1, revision a.1 ed. gaussian, inc. 2009.

[276] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2): 926, 1983.

[277] Abraham M. J., Murtola T., Schulz R., Pall S., Smith J. C., Hess B., and Lindahl E. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1(2):19–25, 2015.

[278] G. A Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi. Plumed 2: New feathers for an old bird. *Comput. Phys. Commun.*, 185(2): 604-613, 2014.

[279] Sousa da Silva Alan W; Vranken Wim F. Acpype - antechamber python parser interface. *BMC research notes*, 2012.

[280] van Zundert G. C. P., Rodrigues J. P. G. L. M., Trellet M., Schmitz C., Kastritis P. L., Karaca E., A. S. J. Melquiond, van Dijk M., de Vries S. J., and Bonvin A. M. J. J. The haddock2.2 web server: User-friendly integrative modeling of biomolecular complexes. *J Mol Biol*, 428(4):720-725, 2016.

[281] Gathiaka S, Liu S, Chiu M et al. D3r grand challenge 2015: Evaluation of protein–ligand pose and affinity predictions. *J Comput Aided Mol Des*, 30(9):651-668, 2016.

[282] Gaieb Z, Liu S and Gathiaka S. D3r grand challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J Comput Aided Mol Des*, 32(1):1-20, 2018.

[283] Gaieb Z, Parks CD, Chiu M et al. D3r grand challenge 3: blind prediction of protein–ligand poses and affinity rankings. *J Comput Aided Mol Des*, 33(1):1-18, 2019.

[284] A. Basciu, P. I. Koukos, G. Malloci, A. M.J.J. Bonvin and A.V. Vargiu. Coupling enhanced sampling of the apo-receptor with template-based ligand conformers selection: performance in pose prediction in the d3r grand challenge 4. *J Comput Aided Mol Des*, Advanced Online Publication, 2019.

[285] Venugopal C, Demos C, Jagannatha Rao K et al. Beta-secretase: Structure, function, and evolution. *CNS Neurol Disord Drug Targets*, 7(3):278-94, 2008.

[286] Nalivaeva NN, Turner AJ. The amyloid precursor protein: A biochemical enigma in brain development, function and disease. *FEBS Lett*, 587(13):2046-54, 2013.

[287] Cole SL, Vassar R. The alzheimer's disease beta-secretase enzyme, bace1. *Mol Neurodegener*, 2(5):22-28, 2007.

[288] Murphy MP, LeVine H. Alzheimer's disease and the amyloid-beta peptide. *J Alzheimers Dis*, 19(1):311-23, 2010.

[289] Prati F, Bottegoni G, Bolognesi ML, Cavalli A. Bace-1 inhibitors: From recent single-target molecules to multitarget compounds for alzheimer's disease: Miniperspective. *J Med Chem*, 61(3):619-637, 2018.

[290] Moussa CE-H. Beta-secretase inhibitors in phase i and phase ii clinical trials for alzheimer's disease. *Expert Opin Investig Drugs*, 26(10):1131-1136, 2017.

[291] Koukos PI, Xue LC, Bonvin AMJJ. Protein-ligand pose and affinity prediction: Lessons from d3r grand challenge 3. *J Comput Aided Mol Des*, 33(1):83-91, 2019.

[292] Weininger D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J Chem Inf Model*, 28(1):31-36, 1988.

[293] Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435-41, 1985.

[294] Peter Willett. The calculation of intermolecular similarity coefficients using an inverted file algorithm. *Analytica Chimica Acta*, 138(1):339-342, 1982.

[295] Wang Y, Backman TWH, Horan K and Girke T. fmcsr: mismatch tolerant maximum common substructure searching in r. *Bioinformatics*, 29(21):2792-4, 2013.

[296] Cao Y, Charisi A, Cheng L-C et al. Chemminer: a compound mining framework for r. *Bioinformatics*, 24(15):1733-4, 2008.

[297] Hong L, Tang J. Flap position of free memapsin 2 (beta-secretase), a model for flap opening in aspartic protease catalysis. *Biochemistry*, 43(16):4689-95, 2004.

[298] Paul C. D. Hawkins, A. Geoffrey Skillman, Gregory L. Warren, Benjamin A. Ellingson and Matthew T. Stahl. Conformer generation with omega: Algorithm and validation using high quality structures from the protein databank and cambridge structural database. *J Chem Inf Model*, 50(4):572-84, 2010.

[299] Hawkins PCD, Skillman AG and Nicholls A. Comparison of shape-matching and docking as virtual screening tools. *J Med Chem*, 50(1):74-82, 2007.

[300] et al Altschul SF, Gish W, Miller W. Basic local alignment search tool. *J Mol Biol*, 215(3):403-10, 1990.

[301] et al Williams CJ, Headd JJ, Moriarty NW. Molprobity: More and better reference data for improved all-atom structure validation. *Protein Sci*, 27(1):293-315, 2019.

[302] et al Case, D.A., Ben-Shalom I.Y., Brozell S.R. Amber18. *University of California*, San Francisco.

[303] Wang L-P, Martinez TJ, Pande VS. Building force fields: An automatic, systematic, and reproducible approach. *J Phys Chem Lett*, 5(11):1885-1891, 2014.

[304] Joung IS, Cheatham TE. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J Phys Chem B*, 112(30):9020-41, 2008.

[305] M. Gerstein and W. Krebs. A database of macromolecular motions. *Nucleic Acids Res*, 15(2): 4280–4290., 1998.

[306] Dewar Michael J. S., Zoebisch Eve G., Healy Eamonn F., Stewart James J. P. Development and use of quantum mechanical molecular models. 76. am1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc*, 107(13):3902-3909, 1985.

[307] Jain AN, Cleves AE, Gao Q, et al. Complex macrocycle exploration: parallel, heuristic, and constraint-based conformer generation using forcegen. *J Comput Aided Mol Des*, 33(6):531-558, 2019.

[308] Cleves AE, Jain AN. Forcegen 3d structure and conformer generation: from small lead-like molecules to macrocyclic drugs. *J Comput Aided Mol Des*, 31(5):419-439, 2017.

[309] Chen I-J, Foloppe N. Tackling the conformational sampling of larger flexible compounds and macrocycles in pharmacology and drug discovery. *Bioorg Med Chem*, 21(24):7898-920, 2013.

[310] P.C. Whitford, O. Miyashita, Y. Levy and J.N. Onuchic. Conformational transitions of adenylate kinase: switching by cracking. *J Mol Biol*, 366(5):1661-71, 2007.

[311] P.C. Whitford, O. Miyashita, Y. Levy and J.N. Onuchic. Conformational transitions of adenylate kinase. *J Biol Chem.*, 283(4):2042-2048, 2008.

[312] L. Rundqvist, J. Adén, T. Sparrman, M. Wallgren, U. Olsson and M. Wolf-Watz. Noncooperative folding of subdomains in adenylate kinase. *Biochemistry*, 48(9):1911-27, 2009.

[313] L. Dechang, S. L. Ming, and Ji Baohua. Mapping the dynamics landscape of conformational transitions in enzyme: The adenylate kinase case. *Biophys J*, 109(3):647-660.

[314] Muller CW, Schlauderer GJ, Reinstein J, Schulz GE. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure*, 4(2):147-156, 1999.

[315] Muller CW, Schulz GE. Structure of the complex between adenylate kinase from escherichia coli and the inhibitor ap5a refined at 1.9 a resolution. a model for a catalytic transition state. *J Mol Biol*, 224(1):159-77, 1992.

[316] Petras Dzeja and Andre Terzic. Adenylate kinase and amp signaling networks: Metabolic monitoring, signal communication and body energy sensing. *Int J Mol Sci*, 10(4): 1729–1772, 2009.

[317] Ahmed A, Rippmann F, Barnickel G, Gohlke H. A normal mode-based geometric simulation approach for exploring biologically relevant conformational transitions in proteins. *J Chem Inf Model*, 51(7):1604-22, 2011.

[318] L. Ponzoni , G. Polles, V. Carnevale, C. Micheletti. Spectrus: A dimensionality reduction approach for identifying dynamical domains in protein complexes from limited structural datasets. *Structure*, 23(8):1516-1525, 2015.

[319] Bakan A., Nevins N., Lakdawala A. S. and Bahar I. Druggability assessment of allosteric proteins by dynamics simulations in the presence of probe molecules. *J. Chem. Theory Comput*, 8(7):2435-2447, 2012.

[320] F. L. Oleinikovas V., Saladino G., Cossins B. P., Gervasio. Understanding cryptic pocket formation in protein targets by enhanced sampling simulations. *J. Am. Chem. Soc*, 138(43):14257-14263, 2016.

[321] S. Banu Bolia, Ashini; Ozkan. Adaptive bp-dock: An induced fit docking approach for full receptor flexibility. *J Chem Inf Model*, 56(4):734–746, 2016.

[322] Leis S., Zacharias M. Efficient inclusion of receptor flexibility in grid-based protein–ligand docking. *J. Comput. Chem*, 32(16):3433-9, 2011.

[323] Pietrucci F., Laio A. A collective variable for the efficient exploration of protein beta-sheet structures: Application to sh3 and gb1. *J. Chem. Theory Comput*, 5(9):2197-201, 2009.

[324] Max C. Watson, Joseph E. Curtis. Probing the average local structure of biomolecules using small-angle scattering and scaling laws. *Biophys J*, 106(11): 2474–2482, 2014.

[325] Thomas A. Halgren. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model*, 49(2):377-89, 2009.

[326] A. T. Laurie and R. M. Jackson. Q-sitefinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics.*, 21(9):1908-16, 2005.

[327] M. Hernandez, D. Ghersi, and R. Sanchez. Sitehound-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res*, 37(Web Server issue):W413-W416, 2009.

[328] B. Huang. Metapocket: A meta approach to improve protein ligand binding site prediction. *OMICS*, 13(4): 325-330, 2009.