Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

SCIENZE BIOTECNOLOGICHE E FARMACEUTICHE

Ciclo XXXII

Settore Concorsuale: 03/D1

Settore Scientifico Disciplinare: CHIM/08

Dynamic Docking, Path Analysis and Free Energy Computation in Protein-Ligand Binding

Presentata da: Martina Bertazzo

Coordinatore Dottorato

Prof. Maria Laura Bolognesi

Supervisore

Prof. Andrea Cavalli

Co-supervisori

Prof. Matteo Masetti
Dr. Sergio Decherchi

Esame finale anno 2020

TABLE OF CONTENTS

ABSTRACT

Comprehending how drugs interact with biological macromolecules to form a complex with consequent biological response is particularly relevant in drug design to guide a rational design of new active compounds. The establishment and the duration of the protein-ligand binding complex is principally determined by thermodynamics and kinetics of the dynamical process of molecular recognition. Thus, an accurate characterization of the free energy governing the formation of the protein-ligand complex is of fundamental importance to deeply understand each contribution to the establishment of the molecular complex.

Experimental biophysical techniques, such as ITC (Isothermal Titration Calorimetry), NMR (Nuclear Magnetic Resonance) and SPR (Surface Plasmon Resonance) proved to be efficient in characterizing both thermodynamics and kinetics of protein-ligand binding. However, a detailed description of the whole binding process on a mechanistic level, with the characterization of all the metastable states is not possible since only a quantitative estimation is allowed.
Conversely, from the computational point of view, plain molecular dynamics (MD), which has been increasingly considered as the method of choice to investigate the entire dynamic process upon complex formation and to predict the associated thermodynamic and kinetic observables, cannot be applied in a routinely drug discovery pipeline because of the high computational cost. In particular, in order to reconstruct a reliable free energy associated to the protein-ligand binding process, all the configurational space has to be extensively sampled, taking longer than the time commonly available in a usual campaign of drug discovery.

In this context, this PhD thesis wants to address specific aspects of the protein-ligand binding process. In particular, it will deal with dynamic docking, thermodynamics and kinetics of protein-ligand binding by devising respectively three different computational protocols. We developed a dynamic docking protocol based on potential-scaled (sMD) simulations, in which both the protein and the ligand are let completely flexible in order to predict the protein-ligand binding pose within a reasonable computational time (Chapter 3.1). Then, we further investigated the applicability of sMD in describing the kinetic behavior of a series of drug-like molecules and we devised a fully automated method to analyze the unbinding trajectories, detecting common features (Chapter 3.2). Finally, we develop a semi-automated protocol based on path collective variables (PCVs) combined with well-tempered metadynamics (well-tempered MetaD) to estimate free energies along a binding path (Chapter 3.3).

LIST OF ACRONYMS

ITC	Isothermal Titration Calorimetry
SPR	Surface Plasmon Resonance
NMR	Nuclear Magnetic Resonance
HTS	High-Throughput Screening
MD	Molecular Dynamics
FES	Free-Energy Surface
MM-PBSA	Molecular Mechanics Poisson-Boltzmann Surface Area
MM-GBSA	Molecular Mechanics Generalized-Born Surface Area
GPU	Graphical Processor Unit
GSK3β	Glycogen Synthase Kinase 3β
HSP90α	Heat Shock Protein 90α
PBC	Periodic Boundary Conditions
PME	Particle-Mesh Ewald
CV	Collective Variable
US	Umbrella Sampling
SMD	Steered Molecular Dynamics
MetaD	Metadynamics
REMD	Replica Exchange Molecular Dynamics
aMD	Accelerated Molecular Dynamics
McMD	Multicanonical Molecular Dynamics
PT	Parallel Tempering
sMD	Potential-scaled Molecular Dynamics
PES	Potential Energy Surface
PCVs	Path Collective Variables
MSD	Mean Square Deviation
RMSD	Root Mean Square Deviation
MSM	Markov State Model
HTMD	High Throughput Molecular Dynamics
SuMD	Supervised Molecular Dynamics
BS-MetaD	Bias-Exchange Metadynamics
RMD	Reconnaissance Metadynamics
SITMD	Selective Integrated-Tempering-Sampling Molecular Dynamics
ENM	Elastic Network Model
VMD	Visual Molecular Dynamics

COM	Center of Mass
RMSIP	Root Mean Square Inner Product
PCA	Principal Component Analysis
MDS	MultiDimensional Scaling method

1 INTRODUCTION

1.1 Molecular Recognition as the basis for drug action

Molecular recognition refers to process in which macromolecules, such as proteins, bind small molecules through noncovalent interactions to form a binary complex. This mechanism, which regulates several biological processes and plays a central role in disease and homeostasis, deals with two important attributes that describe the protein-ligand interaction: specificity and affinity. Specificity is related to the capability of proteins to prefer those binding partners that are highly specific than those that are less specific, while affinity is related to the strength of the protein-ligand complex.[1] Molecular recognition is thus of fundamental importance in the development of drugs, especially during the lead optimization phase, since affinity and specificity determine whether a molecule has the potential to become a drug.

Traditionally, the activity of drug-like molecules has been expressed in terms of the equilibrium dissociation constant, K_d, or by half-maximal inhibitory concentration IC_{50}, both often performed under closed *in vitro* conditions. K_d measured *in vitro,* can be, in certain specific cases, directly correlated to the *in vivo* efficacy of a ligand. However, most frequently, this relation does not directly hold, particularly when the *in vivo* efficacy is mostly determined by the duration of the pharmacological effect. Indeed, sometimes the time for which the biological target is occupied by a ligand, commonly referred to as drug target residence time, t_r, that corresponds to the inverse of the dissociation rate constant ($t_r = 1/k_{off}$), could be more appropriate to describe the efficacy than binding affinity.[2] This aspect recently propelled the need of including the kinetic profiles, expressed in terms of the association rate constant, k_{on}, and the dissociation rate constant, k_{off}, in drug discovery programs, especially during the drug optimization phase.[2] It is therefore of the utmost importance for a deep knowledge of drug action that the processes responsible for the protein-ligand interactions are well characterized mechanistically and quantitatively in energetic terms.[3]

In the first section of this thesis, a brief introduction about the physiochemical mechanisms that control the protein-ligand association is given. In particular, a general description of the most important concepts related to the molecular recognition, such as binding kinetics and thermodynamics, free-energy, entropy-enthalpy compensation are provided. Then a description of the three existing theories, the lock-and-key, induced fit and conformational selection theories, describing the protein-ligand binding are introduced in order to underline the fundamental importance of considering protein flexibility to study biological processes. The final part of this Introduction section is dedicated to experimental and computational methods applied to investigate the protein-ligand binding mechanism.

1.2 Thermodynamics and kinetics of biological complexes

The first attempt to provide a systematic description for the temperature dependence of equilibrium constants of a reaction was proposed in 1889 by Svante Arrhenius. In particular, according to Arrhenius equation, the rate constant k of a chemical reaction can be calculated as:

$$k = Ae^{-\frac{E_a}{k_B T}} \quad (1)$$

Where E_a is the activation energy of the process, T is the temperature, and the pre-exponential term A, is a constant that describes the number of times two molecules collide, and it is characteristic for each chemical reaction since it is a function of the collision radius between two interacting particles and the reduced mass of the system. Despite the Arrhenius equation has been used to determine activation energies for chemical reactions, it can be exploited only to the kinetics of gas reactions in which the transformation to form the product does not involve any intermediate. The transition state theory, as coincidentally formulated in 1935 by Henry Eyring, Meredith Gwynne Evans and Michael Polanyi, was developed to describe more accurately the kinetic behavior of the chemical systems. According to this theory, it is possible to gain insight into rates of reactions by investigating the activated complexes, the so-called transition states. In situations of chemical equilibrium between reactants and activated transition state complexes, the former could convert into products and the kinetic theory could be applied to obtain the rate of this conversion.

According to transition state theory applied to a fully solvated protein-ligand complex, along a single binding path, the dynamics of the transition depends only on the slowest step involved (the one associated to the highest energy). However, reaching an accurate description of the potential energy surface (PES) of the system, especially for the transition states, is far from trivial and, as a result, the absolute reaction rate constants are difficult to obtain.[4]

It is important to underline that the binding or unbinding kinetics is not defined by only one transition state but different pathways are possible, with different transition states associated, can contribute to the observable k_{on} and k_{off} values, so understand which transition state, among all the transition pathways, represents the limiting one could become tricky.[5] Despite these considerations, transition state theory is of fundamental importance in describing protein-ligand interactions and in determining thermodynamics quantities such as Gibbs energy of activation, enthalpy and entropy (Figure 1).

In a first order reaction, the noncovalent association between a protein *P* and a ligand *L* in solution to form the biomolecular complex *PL* can be described as the equilibrium:

$$P + L \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} PL \quad (2)$$

Where k_{on} and k_{off} are the kinetics rate constants that account for both the binding and unbinding reaction. The units of k_{on} and k_{off} are M^{-1}-s^{-1} and s^{-1}, respectively. At equilibrium this two-state mechanism is perfectly symmetrical. In particular the binding reaction is balanced by the reverse unbinding reaction according to:

$$k_{on}[P][L] = k_{off}[PL] \tag{3}$$

Where the equilibrium concentration of all the species involved in the reaction are represented by the square brackets. The binding affinity is then described by the dissociation constant K_d (in unit of M):

$$K_d = \frac{[P][L]}{[PL]} \tag{4}$$

In this situation, K_d corresponds to the ligand concentration at which half of the receptor binding sites are occupied. K_d is directly associated to the free energy difference between the bound and unbound states, ΔG_d. According to equation (3) and equation (4), the thermodynamics constant K_d is linked to the kinetics constants k_{on}, k_{off}, as follow:

$$K_d = \frac{k_{off}}{k_{on}} \tag{5}$$

Therefore, it is interesting to understand how complexes with the same affinity could have, at least in principle, very different transition states. Indeed, varying the transition state decreases or increases both kinetic rates without altering the energy difference between both the bound and unbound states. This because the rate constants k_{on} and k_{off} depend on the transition state along the binding pathway that separates the bound and unbound states:

$$k_{on/off} = \frac{k_B T}{h} e^{\frac{\Delta G^{\ddagger}_{on/off}}{RT}} \tag{6}$$

According to the above equation, the kinetic constants, k_{on}, k_{off}, are proportional to the exponential of the energies of activation of the respective transitions state, $\Delta G^{\ddagger}_{on/off}$, through a pre-exponential factor that combines the Boltzmann's constant k_B, the Planck's constant h , and the absolute temperature T.

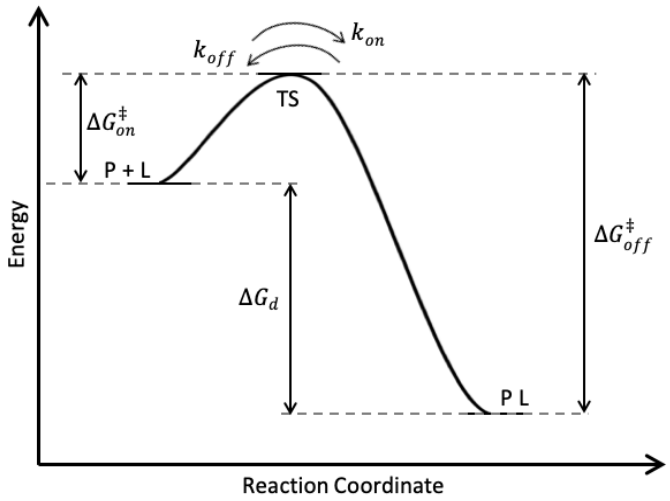


Figure 1. Simplified free-energy profile of the protein-ligand complex PL formation between a protein P and a ligand L. TS represents the transition state. The thermodynamics of binding is quantified by the free energy difference between both the bound and unbound states (ΔG_d), and the kinetics is determined by the dissociation and the association rate constants, k_{on}, k_{off}. These quantities are related to the free-energy differences between minima and the transition state TS, $\Delta G^{\ddagger}_{off}$ and $\Delta G^{\ddagger}_{on}$, respectively.

As already discussed, much consideration has been given to the impact of including kinetics information in the optimization phase of the drug discovery process, since models based both on kinetic constants and binding affinity can be more reliable.

The protein-ligand binding process in a solvent is an example of thermodynamic system so it is possible to describe the driving forces that dictate the association of the two entities using the laws of thermodynamics. In particular, as any spontaneous process, the establishment of the protein-ligand complex PL occurs only when the change in Gibbs free energy, ΔG_d, of the system is negative, once the system reaches an equilibrium state at constant pressure and temperature. At the steady state, ΔG_d can also be defined as the variation of the unbound to the bound state of two state functions of the system: enthalpy (H) and entropy (S) with the following equation:

$$\Delta G_d = \Delta H - T\Delta S \tag{7}$$

For a binding process, the binding enthalpy, ΔH, reflects the energy change of the system when the ligand binds to the protein, resulting from the formation of noncovalent interactions between them.[6] Change in binding enthalpy is a global propriety of the entire system. This means that the change upon biding involves also contributions from the solvent, in particular the formation of new interactions between the ligand and the protein coincides with the disruption of interactions between ligand and solvent, protein and solvent and a new solvent reorganization near the complex

surface.[1] Entropy is a measure of the disorder or randomness of the overall system (including the ligand, protein and surrounding solvent). ΔS is a global thermodynamic observable of a system which is positive when the overall degree of freedom of the system increase and, on the contrary, the negative sign indicates decrease in degrees of freedom of the system. The binding entropy, ΔS, is determined by:

$$\Delta S = \Delta S_{solv} + \Delta S_{conf} + \Delta S_{r/t} \quad (8)$$

Where ΔS_{solv} reflects the solvent entropy change in the accessible surface area of the protein and ligand upon complexation; ΔS_{conf} represents changes in the conformational entropy reflecting the changes in the conformational degree of freedom of protein and ligand upon binding and $\Delta S_{r/t}$ represents the loss of rotational-translational degrees of freedom of protein and ligand upon binding.[7]

As introduced before, the sign and magnitude of ΔG_d are determined by the variation of entropy and enthalpy. Thus, ΔH and ΔS can be considered as the driving forces for the protein-ligand binding event. In particular, negative enthalpy change, which is associated with the establishment of new noncovalent interactions between association partners, is accompanied by a negative entropy change caused by the loss of degrees of freedom of protein and ligand. This phenomenon, where variation of ΔG_d is caused by mutual variation in ΔH and ΔS, is called the entropy-enthalpy compensation.[8] Different mechanisms can contribute to the entropy-enthalpy compensation, including the structural and thermodynamic proprieties of the solvent, the protein flexibility and the molecular structure of the ligand.

1.3 Protein-ligand binding models

The mechanisms underlying biomolecular recognition have been deeply investigated, in order to understand these processes and define new active compounds in a drug discovery context.[9–13] Three different models have been introduced for ligand recognition, the "lock-and-key",[9] "induced fit"[11] and "conformational selection"[13–16] (Figure 2).

According to the lock-and-key model (Figure 2A), the protein and the ligand are considered rigid and they possess specific complementary shapes that match exactly into one another, like a key into a lock, with almost no change in protein conformation. The binding event is based on the grounds that desolvation of the protein and the ligand during the process leads to changes in entropy contribution. In particular, the collision between the protein and the ligand causes a complete displacement of the water network surrounding the protein and ligand interaction interfaces,

causing a positive enthalpy change, while the release of water increases the solvent entropy. According to this model, that considers the protein as rigid, there is no change in conformational entropy. Therefore, the solvent entropy should be large enough to compensate the positive enthalpy change due to the desolvation process and also the negative entropy caused by the loss of rotational and translational motions of the ligand. The lock-and-key model fails to explain the experimental evidence that a protein is also able to bind to a ligand even in the case where their initial shapes do not fit perfectly.

Conversely, the induced fit model, considering the flexibility of both the interacting binding partners, the ligand and the protein binding site, is able to provide a plausible explanation (Figure 2B). According to induced fit model, during the interaction with the ligand, the protein undergoes a conformational change only in the binding site, neglecting major conformational changes that could involve the whole protein structure. The absence of the initial perfect surface complementary between protein and ligand results in multiple tentative collisions to establish favorable contacts (negative enthalpy change), that should be strong enough to provide the encounter complex enough strength so that induced fit takes place in a reasonable time.[1] Particularly, in order to maintain the stability of the protein-ligand complex, the negative enthalpy change contribution, resulting from the establishment of interactions between protein and ligand should be greater than the sum of the positive enthalpy change resulting from the disruption of the interaction with the solvent, the negative entropy change resulting from reducing the conformational freedoms of the interacting surfaces, ΔS_{conf}, and the rotational-translational degrees of freedom of the binding partners, $\Delta S_{r/t}$.

Furthermore, the lock-and-key and induced fit models consider the protein, under given experimental conditions, as a single and stable conformation. Instead, most proteins are intrinsically dynamic, and the conformational selection model considers their flexibility (Figure 2C). In accordance with this model, the native state of a protein exists as an ensemble of different conformations, all coexisting in equilibrium with different population distributions, and the ligand binds the one that is more selectively altering the equilibrium toward this state. It is difficult for the conformational selection model to understand which state of the system, entropy or enthalpy, contributes the most to the lowering of the system's free energy. In particular, it seems like the selective binding of the ligand to a specific conformational state of the protein is dominated by the solvent entropy gain due to the desolvation effect (as occurs in lock-and-key-model), while, the following conformational adjustments of the protein is dominated by the system enthalpy decrease due to the formation of newly interactions between interacting partners.

It is important to underline that, because these three conceptual models for biomolecular recognition have been observed experimentally, they may exist simultaneously or in a sequential

manner, and also that even more complicated mechanisms than those presented here may be possible.[17]

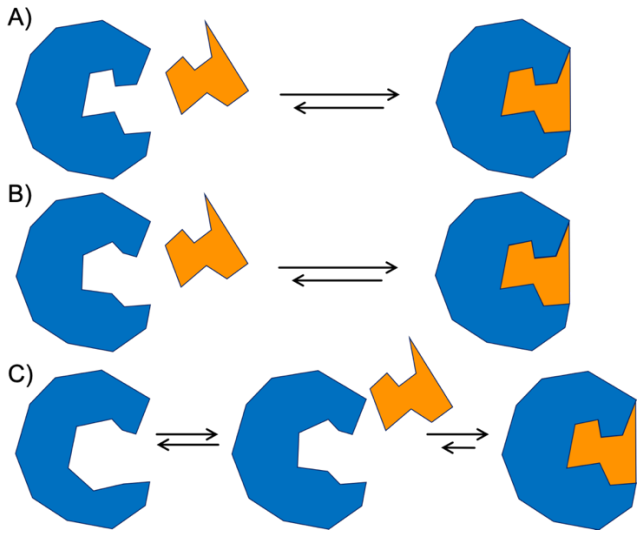


Figure 2. Schematic illustrations of the three different protein-ligand binding models: the lock-and-key (A), the induced fit (B), and the conformational selection (C). The protein is represented in blue while the ligand is represented in orange.

1.4 Experimental and computational strategies to characterize protein-ligand interactions

Several experimental and computational techniques could be used to investigate different aspects of protein-ligand binding. In this regard, isothermal titration calorimetry (ITC), nuclear magnetic resonance (NMR) and surface plasmon resonance (SPR) are briefly introduced in the first part of this paragraph as experimental methods to characterize protein-ligand binding while, in the second part, molecular docking and binding free energy calculations are discussed as theoretical approaches.

1.4.1 Experimental methods

ITC is a biophysical technique that allows a direct estimation of the heat exchange during the complex formation at constant temperature; this is probably the gold standard in determining the energies driving the binding process.[18] ITC allows simultaneous determination of binding affinity K_d, stoichiometry n, enthalpy change ΔH, calculation of free energy change ΔG_d and entropy

change ΔS in one single experiment. A typical ITC experiment is a three steps procedure. In the first step, the ligand is titrated into a solution containing the biomolecular target of interest.
The second step consists in measuring the heat absorbed or released that is associated with the binding event. In particular, the temperature imbalance between the reference and sample cells, due to protein-ligand binding event, is measured and compensated by modulating the feedback power applied to the cell heater. During the last step the primary ITC data, that are the power applied to the sample cell as a function of time, are processed and fit to obtain the binding curve representing the heat of reaction per injection as a function of the ratio of the total ligand concentration to the protein concentration. Finally fitting the binding curve, the binding constant $K_b = 1/K_d$, the binding enthalpy ΔH, and the stoichiometry of the binding event n are obtained. Consequently, knowing the binding constant, the standard Gibbs free energy ΔG_d, the binding entropy ΔS can be derived. The heat exchange revealed by ITC is the total heat effect in the sample cell consequent the ligand addition, including the heat adsorbed or released during the binding event as well as the heat effects due to the dilution of the ligand and protein, the mixing of two different solutions, the different temperature between the sample and references cells, and so on. Therefore, evaluating the heat change due to the contribution of binding only is far from trivial.

NMR spectroscopy is a particularly efficient method exploited to get information about protein-ligand interactions at atomic resolution.[19] Several NMR spectroscopy approaches exist to study protein-ligand complex which can be categorized into two general group: protein observed or ligand observed techniques.
In protein observed methods, a spectrum of protein is obtained, and the ligand is titrated, thus providing information about residues of the protein directly involved in the interaction with the ligand. Conversely, in ligand observed techniques, the spectrum of ligand is obtained, and the protein is added. The overall NMR spectra depends on the lifetime of the protein-ligand complex which is given by the residence time (t_r). In particular, for strong complexes the lifetime of the protein-ligand complex is much longer than the difference in chemical shifts between the signals obtained for the unbound and bound form, thus two separate NMR signals are obtained. On the contrary, when weak complexes are involved, the lifetime of the complex is too short to observe the two signals independently and a single NMR signal is obtained.

SPR spectroscopy[20] is a popular technique used for the estimation of association and dissociation rate constants during protein-ligand binding or unbinding events. SPR is an optical-based method that assesses the change in the refractive index near the sensor surface. It is a label-free technique, which is advantageous in comparison with the radioligand binding assays that have been previously used for the biochemical characterization of the formation of specific drug-target complexes. In the

most widespread configuration, the sensor surface is a thin gold film on a glass support, which is positioned on the bottom of the flow cell through which an aqueous solution flow.
The receptor molecules are immobilized on the sensor surface and the ligand (analyte in SPR terminology), is injected into the aqueous solution. As the analyte binds to the immobilized receptor, an increase in the refractive index is observed. Once all the binding sites are occupied, a running buffer without analyte is injected through the flow cell to let the ligand molecules dissociate from the target protein. As the analyte dissociates, a decrease in the refractive index is measured. The time-dependent resonance unit (RU) curve is processed and fitted to determine the association and dissociation rates, k_{on} and k_{off}. The equilibrium dissociation constant K_d is quantified by fitting the resonance unit sinusoidal curve as a function of the analyte concentration. Moreover, the binding enthalpy can be estimated by van't Hoff analysis.[21] It is important to note that by immobilizing the protein, the conformational and roto-translational entropies may be affected impacting on the evaluation of the association rate constants.

1.4.2 Computational methods

Even if experimental techniques can investigate thermodynamic profiles for a ligand-protein complex, the operations for measuring the binding affinity are laborious, time-consuming, and expensive. In a modern rational drug design campaign, to find a set of lead molecules, high-throughput screening (HTS) of a large library of hundreds or thousands of samples is involved. Thus, it is hardly feasible or even possible to achieve this goal using only experimental methods. The opportunity to predict completely *in silico* ligand binding modes and to investigate binding processes to identify and optimize new lead candidates motivates the Scientific Community to put great efforts in developing new computational techniques. Molecular docking can predict, for instance, very quickly which ligand fits best into the binding pocket of the protein target and assess the binding affinity of the complex. More accurate predictions of binding affinity could be obtained through molecular dynamics simulations, which take in account all thermodynamically relevant phenomena such as the protein flexibility and explicit inclusion of the solvent.

1.4.2.1 Molecular Docking

Molecular docking is a well-established computational strategy for predicting *in silico* the binding modes and affinities of molecular recognition events.[22] Thank to the constant growth of available protein structures and to the improvement in computational resources, molecular docking has been proven to be an important methodology for drug discovery campaigns.[23] This procedure allows to dock compounds rapidly, allowing the selection for moving forward biological tests of those

compounds that are likely to have an action against the target of interest, thus saving time and costs. However, this computational speed occurs at the cost of accuracy, especially when protein rearrangements are required upon ligand binding.[24,25] Indeed, protein flexibility is not treated extensively; additionally water molecules, which can be crucial for reproducing the specific protein-ligand complexes, are very difficult to treat explicitly. However, thanks to the methodology advances several different software have been developed to include explicit water molecules into docking and virtual screening studies[26] such as: GRID,[27] HINT,[28] Superstar,[29] JAWM,[30] WaterMap,[31] Water PMF[32] and Water FLAP.[33] Therefore, an accurate evaluation of the thermodynamic and kinetic quantities is not allowed.

At the time of the earliest implementations of molecular docking, around 1982 when the first molecular docking was reported by Kunts et al.[22], both ligand and protein were treated as rigid bodies but, nowadays, full flexibility of the ligand and partial flexibility of the receptor are implemented by almost all docking programs.[34]

A typical docking algorithm consists of two main components: the search algorithm and the scoring function. The former component, which can be either stochastic or deterministic, is responsible for searching through different ligand conformations and orientations (poses) within a given target receptor's binding site. As already introduced, the early docking implementations dealt with a unique rigid conformation of the protein due to limited computational resources. However, proteins are intrinsically dynamic and, most of times, there is a mutual adaption between protein and ligand in order to maximize favorable interactions upon binding.[35] In principle, according to the more recent interpretation of molecular recognition models, the degrees of freedom of protein and ligand should be sampled simultaneously, but this is still too computational demanding, so a number of different strategies were developed to provide a solution.

These can be split into two main groups: single-structure methods or ensemble methods. Single-structure methods are related to the induced-fit model in which only the binding pocket is perturbed by on-the-fly local changes after ligand binding conformational search. Alternatively, in ensemble methods, an ensemble of previously generated conformations is exploited in a series of independent docking procedures mimicking the conformational selection model. Anyway, information provided by induced-fit approaches and multiple protein conformations is difficult to manage and it must be properly used to be effectively exploited.[36,37]

Scoring functions are approximate mathematical methods used to assess the binding affinity (generally through estimating the strength of noncovalent interactions) in a protein ligand complex. In order to reduce the algorithm's complexity, since many different physical interactions (including those with the solvent) as well as the entropic effects need to be considered in a reliable estimation of binding free energy, a number of simplifications are introduced at the cost of accuracy.[38] The most important scoring functions available for protein-ligand docking can be divided in three

general classes: the force-field-based, the empirical-based and the knowledge-based (also known as statistical potentials) scoring functions.

In the force-field-based approaches, binding affinities are estimated by using functional forms and parameters, defined as force-fields, derived from experiments and quantum mechanical calculations. In order to reduce the complexity, usually only the enthalpic contribution, mostly due to the strengths of intramolecular noncovalent interactions between interacting partners, is taken into account. The desolvation energies of the ligand and of the protein are sometimes re-estimated using implicit (or continuum) solvation methods such as Molecular Mechanics Poisson-Boltzmann surface area (MM-PBSA)[39,40] and the generalized-Born surface area version (MM-GBSA).[41,42] Empirical scoring functions are based on the parametrization of different types of interactions, usually via regression methods, as favorable or unfavorable energy terms contributions to the binding affinity.[43] The knowledge-based scoring functions assume that statistical observations of intramolecular close contacts between protein and ligand that occur more frequently than those expected by a random distribution are likely to be energetically favorable and therefore contribute favorably to binding affinity. Thus, these contacts are used to derive the statistical potentials. All these three scoring functions, known as classical scoring functions, assume an additive functional form to represent the linear relationship between the binding affinity and those features that describe the protein-ligand complex. By contrast, Machine learning scoring functions, which use pattern recognition algorithms to discriminate mathematical relationships between empirical observations, do not make suppositions about the form of the functional.[44,45]

In summary, molecular docking represents a valid method to identify crystallographic binding modes and to select new promising compounds from entire libraries of molecules, especially when protein flexibility does not play a fundamental role in the binding process. Instead, when large rearrangements are involved in the protein-ligand binding process, molecular docking suffers from severe limitations. Moreover, molecular docking does not provide the necessary accuracy to calculate a reliable estimation of binding free energy and it cannot be used to estimate kinetics quantities. In principle, all these problems could be overwhelm by molecular dynamics (MD) simulations and related methods.[46]

1.4.2.2 Molecular Dynamics (MD) simulations

MD simulations describe the physical evolution of the system in time solving Newton's equations of motion, immersed in a thermic bath, where forces between the interacting particles and their potential energies are calculated using molecular mechanics force fields. During the last decade, thanks to the advent of faster architectures and development of more specific computation algorithms, MD simulations have become a valid alternative to traditional molecular docking in

drug design since the exploration of the entire protein-ligand process at full atomistic level is allowed, in principle. Moreover, the possibility to investigate the entire protein-ligand binding process could shed light on metastable states sampled by the ligand, protein conformational rearrangements during or consecutive the ligand binding, alternative binding sites, and also the contribution of the water during binding.[47] Additionally, provided that simulations of the binding process are long enough to observe the entire mechanism, from the drug completely solvated water (unbound state) to the formation of protein-ligand complex (bound state), thermodynamics and kinetics observables can be accurately evaluated through the estimation of the free-energy surface (FES) as described in the following section of this thesis (2.2 Estimation of MD-derived observables).[46,48] However, multiple binding events have to be observed along the simulation in order to collect an adequate statistics to correctly calculate thermodynamics and kinetic quantities. This usually require milliseconds simulations making the process too computationally expensive to be routinely used instead of traditional molecular docking methods especially in the first part, when novel compounds have to identified or during the hit-to-lead optimization.

In order to reduce the computational cost required, a number of different enhanced sampling techniques have been developed during the last years. These methods accelerate MD simulations, increasing the probability of observing binding process, by applying biasing forces or altering the potential energy function. Thanks to these techniques combined with the ever-increasing computational power, MD simulations are gaining everyday more attention in drug discovery process, replacing the standard molecular docking with the new concept commonly referred as "dynamic docking".[25,48]

1.5 Summary of the contribution

MD has been increasingly considered as a promising computational technique to study protein ligand-binding since the entire mechanism can be investigated and, in principle, the associated thermodynamic and kinetic observables can be estimated. However, due to the significant computational effort involved, the direct application of plain MD to study protein-ligand binding is often not possible or highly inefficient, at least from a drug discovery perspective.

In this contribution, we explored possible different protocols, depending on the specific issue to address, based on enhanced sampling MD simulations to gain insights into thermodynamics and kinetics associated to the protein-ligand binding process.

In particular, three different aspects of protein-ligand binding are considered: prediction of ligand binding mode into the target binding site, estimation of residence time for a congeneric series of

compounds binding to the same biological target and the computation of the FES of the entire mechanism of ligand binding.

In the first application, a semi-automated protocol based on an enhanced sampling technique, the scaled-potential MD (sMD), is developed in order to predict protein-ligand binding poses. This protocol was validated on its ability to reproduce the crystal binding poses of a set of five different compounds binding two relevant pharmacological targets (the Glycogen Synthase Kinase 3β, GSK3β, and the N-terminal domain of Heat Shock Protein 90α, HSP90α).

As for the second test case, we exploited sMD to perform a series of unbinding simulations of a series of congeneric compounds, inhibitors of HSP90α, in order to prioritize them according to their average computational exit times. Moreover, we devised a protocol based on dimensionality reduction and clustering analysis to qualitatively compare pathway followed by the ligands during unbinding simulations.

Finally, in the last test case, we devised a semi-automated strategy to compute a guess path along with the FES of the whole protein-ligand binding process. To achieve this, we took advantage of a machine learning algorithm introduced by Ferrarotti et al,[49] known as principal path to identify a binding path from pre-existing MD-binding simulations. In addition, we reconstructed the FES exploiting the well-known well-tempered Metadynamics algorithm.

2 THEORETICAL AND METHODOLOGICAL BACKGROUNDS

During the last decades, MD simulations have been optimized properly in order to capture accurately the full protein-ligand binding process in agreement with experimental results. This is due to increasing computer power, the advent of graphical processor unit (GPU) architectures, and MD software that can efficiently run on these innovative infrastructures.[50] As already discussed in Introduction, in pioneering studies by Buch et al.[51] , Shan et al.[52] and Dror et al.[53], several spontaneous protein-ligand binding events reproducing the complexes resolved by X-ray crystallography, were observed. However, as it was previously discussed, several different trajectories are needed to obtain adequate statistics and to sample exhaustively the conformational space. This makes the whole process computationally demanding and not frequently affordable, especially in fast-paced drug discovery, when MD simulations are used for kinetic prediction of ligands with long residence time.[46] However, powerful alternatives to the plain MD simulations have been developed in order to accelerate the observations of rare events on affordable timescales.[54] Most of these approaches, usually referred as "enhanced sampling methods", can provide also the statistical weights of the sampled configuration, allowing the reconstruction of the thermodynamics of the system consistent with the one obtained by a plain MD simulation.
In this section, we provide some theoretical background on methods and combination thereof we employed to study protein-ligand binding process and related thermodynamics and kinetics proprieties of biologically relevant systems.

2.1 Molecular Dynamics (MD) simulations

Molecular dynamics, MD, is a computational technique aimed at studying the time-dependent evolution of a molecular system, under the action of the forces generated by a potential that provides opportune approximations of the physics and chemistry under investigation. This is obtained by solving the second-order differential equations represented by Newton's second law:

$$f_i(t) = m_i a_i(t) = -\frac{\partial V(x(t))}{\partial x_i(t)} \quad (9)$$

Where $f_i(t)$ is the force acting on *i*th atom of mass m_i, at time *t*. $a_i(t)$ is the acceleration and *x(t)* represents a configuration of the system described by three-dimensional coordinates of the *N* interacting atoms at time *t.* Commonly, in computational drug discovery, molecule's atoms are treated as a sphere particles according to molecular mechanics principle. This is possible through the Born-Oppenheimer approximation.[55] According to this formalism, the most important

contribution to the system dynamics is associated to nuclear motions, which are considerably heavier than electrons in term of mass. This allows to average out the electronic contribution in Schröedinger's wave function equation that describes the system. The result is that an empirical potential energy function is introduced $V(x(t))$ that represents the Force Field (FF):

$$V = \sum_i^{bonds} \frac{k_{l,i}}{2}(l_i - l_o)^2 + \sum_i^{angles} \frac{k_{\alpha,i}}{2}(\alpha_i - \alpha_{0,i})^2 + \sum_i^{torsions} \left\{ \sum_k^M \frac{V_{ik}}{2} \left[1 + \cos\left(n_{ik} \cdot \theta_{ik} - \theta_{0,ik}\right) \right] \right\} + \sum_{i,j}^{pairs} \varepsilon_{ij} \left[\left(\frac{r_{0,ij}}{r_{ij}} \right)^{12} - \left(\frac{r_{0,ij}}{r_{ij}} \right)^{6} \right] + \sum_{i,j}^{pairs} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \quad (10)$$

The FF accounts for bonded and non-bonded contributions. In equation (10) the first three terms represent bonded terms describing variations in potential energies as a function of bond stretching, bending and torsion among atoms directly involved in bonding relationships. Bond stretching and bending, represented by summation over bond lengths (l) and angles (α) respectively, are both described by harmonic potentials with reference values l_0 and α_0 and force constants of k_l and k_w. Torsions of dihedral angles are described by a cosine series of M terms, and other parameters are additionally considered that take in account the periodicity. In particular, n_{ik} describes the multiplicity for the kth term of the series, $\theta_{0,ik}$ is the corresponding phase angle, and V_{ik} is the energy barrier. Non-bonded terms describe forces acting among atoms that are not directly bound and consist of van der Waals and Coulomb interactions (fourth and fifth term in equation (10) respectively). These forces vary with the inverse power of the distance between considered atoms r_{ij}. In particular, van der Waals interactions are generally described by a 12-6 Lennard-Jones potential, where ε_{ij} defines the depth of the energy well, whereas $r_{0,ij}$ is the minimum energy distance that equals the sum of the van der Waals radii of the two interacting atoms considered. The last term in equation (10) describes the electrostatic energy that is defined by the Coulomb potential. q_i and q_j are the partial charges of the pair of considered atoms, ε_0 is the permittivity of free space and, finally, ε_r is the dielectric constant (1 in vacuum). The empirical determination of reference values for bond lengths, angles, dihedral together with the different degree of accuracy used to describe the non-bonded interactions limits the applicability of the FF. However, multiple FF with constantly increasing accuracy are now available, so one can always take in account the FF best fitting to the specific considered molecular system. In a normal MD simulation, given a certain FF and potential $V(x(t))$, it is possible to derive the acceleration a over the atoms with respect to positions x at time t. Even if a classical FF speeds up calculations compared to quantum mechanics (QM), equation (10) still it requires integration over time. Thus, to simulate biological macromolecules it is necessary to split the integration of the equation of motion into discrete time intervals, known as time-steps (δt), where forces are assumed to be constant. As an example of an integrator, the velocity-Verlet is one of the most widely adopted algorithm in MD simulations. The Velocity-Verlet integrator calculates positions at time $(t + \delta t)$ through the following equation:

$$x_i(t + \delta t) = x_i(t) + v_i(t)\delta t + \frac{1}{2}a_i(t) \quad (11)$$

Where $x_i(t)$, $v_i(t)$ are the coordinates and velocities at time *t* respectively, accelerations $a_i(t)$ are calculated taking the first derivative of the potential energy with respect to positions with opposite sign as shown in equation (11). Velocities are then calculated according to equation:

$$v_i(t + \delta t) = v_i(t) + \frac{1}{2}[a_i(t) + a_i(t + \delta t)]\delta t \quad (12)$$

The accuracy of integrators heavily depends on the value of the chosen time-step, that must be chosen small enough to track the dynamics of the system.

Only a "small" δt guarantees reliable forces over time and consequently, continuous trajectories. Ideally, the value of the time-step should be comparable to the time scale of the fastest motion among those that are being studied. For example, stretching and bending of bonds involving hydrogen atoms are subjected to very fast motions so a correct value of time-step should be 0.1 fs. This time step is very small and a strategy to increase it is to constrain the fastest degrees of freedom subjected to very fast motions thanks to ad-hoc algorithms (i.e., stretching and bending of bonds involving the hydrogen atoms).[56] In this way, it is possible to neglect these motions and thus consider a time-step 10 times larger, heavily reducing the computational cost. Other strategies are possible: for example, implementing a protocol to repartition the hydrogens mass over adjacently bound heavy atoms, allowing a time step of 4 fs.[57,58] Alternatively these masses are not considered at all and replaced with virtual sites without mass,[58] whose positions are updated relatively to heavy atoms at each step.

An important aspect that any numerical integrator must satisfy when used in an MD simulation, is the preservation of the energy conservation law. In particular, the integrator must keep the Hamiltonian of the system constant, given by the sum of the potential and kinetic energy. According to the following equation:

$$H(x, p) = V(x) + K(p) \quad (13)$$

the Hamiltonian $H(x, p)$ provides the total energy of the system and depends upon coordinates *(x)* and momenta *(p)*. From these considerations, because the macroscopic variables that define the statistical NVE ensemble are considered constant, MD naturally follows the motion of a microscopic isolated system, without exchanging energy with the surroundings. Most of the MD simulations are performed within the canonical ensemble, where the number of particles (N), volume (V) and temperature (T) are conserved, or within the isothermal-isobaric ensemble, where N, V and pressure (P) are conserved. Constant temperature is maintained through a thermostat.[59] Thermostats are algorithms that allow fluctuations in kinetic energy as if the simulated system were

immersed in a thermostatic bath. Constant pressure is controlled by the so-called barostat algorithms,[60,61] that is by opportunely scaling the system volume. In order to limit the finite size effects due to the use of a simulation box, periodic boundary conditions (PBC) are also commonly used in MD simulations. With the employment of PBC conditions, the system is placed in a unit cell that is replicated in different directions resulting in a periodic lattice of identical subunits. In this way, molecules of the system that are close to the edge interact with the atoms of the neighboring box. In order to limit the computational cost of the non-bonded terms calculations, a spherical cut-off with a radius usually at least 11 Angstrom has to be used. In particular, the forces are divided into short- and long-range contributions depending on the distance value as below or over the above mentioned threshold.
The long-ranged electrostatic interactions, because they decay as r^{-1}, are usually calculated over the a lattice through particle-mesh Ewald (PME) method.[62] While short-range electrostatics is computed in the real-space, a grid-based approach is used in the reciprocal space to calculate long-range electrostatic taking advantage of fast Fourier transform.[62]

Generally, MD simulation setup consists of three important steps: minimization, equilibration and production. During the minimization step, the initial structure of the system is relaxed to the nearest local minimum (potential energy).[61] During the equilibration step, the system is brought at physiological temperature and pressure conditions. The initial velocities are assigned to the atoms according to a Maxwell-Boltzmann probability distribution at the desired temperature. This step should be gentle enough to avoid artifacts in the protein structure, so generally, it is a good practice to run few steps in the NVT ensemble, in the first heating steps. Once the temperature reaches an acceptable value (e.g. 300 K), a switch to the NPT ensemble is necessary to relax the system.
By use of these methods, MD simulations on systems of relevant pharmaceutical interest can be performed from the microsecond to millisecond time scale in order to observe events such as drug binding to its target and to estimate the kinetics and free energy associated with protein ligand binding.

2.2 Estimation of MD-derived observables

According to statistical mechanics principles it is possible to quantitively estimate important thermodynamics observables computed from MD runs, such as internal energy, pressure and heat capacity. Among these, as already discussed in the Introduction, the key thermodynamic quantity in drug discovery is the protein-ligand standard binding free energy,[46] ΔG_d°, defined as the free energy difference between the unbound and bound states.[63] ΔG_d° can be directly compared to the experimental dissociation constant K_d through the following equation:

$$\Delta G_d^\circ = k_B T ln\left(\frac{K_d}{C^\circ}\right) \quad (14)$$

Where C° is a constant defining the standard concentration (1 M by convention) that makes the argument of the logarithm dimensionless. [64] Because the standard binding free energy depends on the dissociation equilibrium constant K_d and the reference value of C°, this conventional concentration must be properly considered to compare computational free-energy estimation with experimental data.[64] Together with other quantities related to entropy, free-energy is the thermodynamic observable whose estimation suffers the most from the limitations and underlying approximations of the sampling. Because of the probability of visiting microstates in the canonical ensemble is proportional to the Boltzmann factor according to:

$$p(x) \propto e^{-V(x)/(k_B T)} \quad (15)$$

It is evident that thigh-in-energy configurations are much less frequently sampled than low-energy states. In addition, if barriers larger than $k_B T$ units need to be crossed, the sampling of the configurational space will be severely limited, which is a prerequisite to achieve reliable free-energy estimate. For this reason, moving across different states, separated by significant free-energy barriers, can be viewed as a rare event relative to the time scales accessible to plain MD simulations. Equation (15) can be expressed as a function of the probability rate between bound and unbound states as:

$$\Delta G_d^\circ = -k_B T ln\left(\frac{p_{bound}(x)}{p_{unbound}(x)}\right) + k_B T ln\left(\frac{C^{box}}{C^\circ}\right) \quad (16)$$

Where $p_{bound}(x)$ and $p_{unbound}(x)$ are the probability of finding the protein in the bound or unbound states, respectively. The second term is a correction term to obtain the standard binding free energy. This term considers as well as the reference concentration of the standard state, C° and the concentration of the interacting species included in the simulation box C^{box}. The MD approach, together with Monte Carlo methods, are the most accepted techniques to generate a Boltzmann distributed set of configurations, finding application in studying biomolecular systems. However, assessing free energy as a probability ratio according to equation (16) is far from trivial, since many re-crossing events between the bound and unbound states, separated by energy barriers larger than $k_B T$ units, are required to achieve proper statistical confidence. For plain MD simulation, this is even harder since it is already pretty rare to observe a single binding event in a reasonable computational time.

As it was already widely discussed in Introduction chapter, the main advantage of MD-based methods is the possibility of characterizing the mechanistic steps of the entire protein-ligand binding process.

In particular, provided that adequate sampling is obtained, kinetic observables such as the association and dissociation constants (k_{on} and k_{off}) can also be determined by using equations (5-6).

During the last decade, kinetics received increasing attention in the drug discovery field, following the perspective article of Copeland published in 2006[2] where the concept of residence time define as the reciprocal of the k_{off} is thoroughly discussed:

$$t_r = \frac{1}{k_{off}} \tag{17}$$

His discussion starts from the evidence that t_r is related to the in vivo biological effects triggered by the ligand. In other words, the efficiency of a drug is explicated only when the drug is bound to its physiological target, whose cellular function is consequently modified.[65,66] Thus, from a pharmaceutically standpoint, accurately predicting the t_r can be more important than the determining the equilibrium constant to predict the final in vivo drug efficacy. Therefore, optimizing t_r via rational design is of fundamental importance for computational drug discovery. However, plain MD simulations are not appropriate to recover transition rates for drugs; this is due to the long timescales involved in the dissociation of protein-ligand complexes, which can last from millisecond to minutes or more.

2.3 Enhanced sampling methods

Thanks to improvements in hardware and software, it is now possible to perform long MD simulations able to observe in detail the full protein-ligand binding process, at least in some cases. However, due to the huge time scale needed to adequately sample the configurational space involving the binding event, it is not possible to routinely exploit plain MD in the drug discovery process, making the plain MD binding more like academic exercise. Indeed, most of the times, simulating the whole process of ligand binding to its target requires the exploration of a large number of degrees of freedom, and also the ligand may spend a lot of time in irrelevant regions in the bulk solvent before finding the way to the binding site. As result, the system gets trapped in a local minimum of the FES, without being able to overcome the energy barrier and explore other regions, including that associate with the absolute minimum corresponding to the protein-ligand crystallographic complex.

To overcome this limitation, a number of strategies, the so-called "enhanced sampling methods", have been developed.[46,67] As suggested by the name, they improve the exploration of the configurational space.[67] Broadly speaking, these methods usually fall in two general categories. On one side, one has methods that rely on collective variables (CVs) biasing. The main advantage of using these strategies is that the sampling is enhanced toward the event of interest by biasing the MD simulations along chosen CVs, that are functions of atomic coordinates. Among these methods, the most well-known are umbrella sampling[68] (US), steered MD[69] (SMD), and metadynamics[70] (MetaD). On the other side, one has tempering methods, in which sampling is enhanced along all of the degrees of freedom of the system. Replica exchange MD[71] (REMD), accelerated MD[72] (aMD), multicanonical MD[73] (McMD), Parallel tempering (PT),[71] and potential-scaled MD[74,75] (sMD) belong to this second class.

In this work, we took advantage of both classes of enhanced sampling methods so, in the following, a brief description of the theory behind the employed techniques is provided.

2.3.1 Potential-Scaled MD (sMD)

sMD belongs to the non-CV based class of methods.[74,75] In this method the PES of the system is scaled by a factor λ that belongs to the interval (0,1]. When

λ equals to 1, then the PES corresponds to that obtained from a plain MD. As one goes to lower values of λ, approaching the 0 value, the energy profile is increasingly smoothed. In Figure 3 it is depicted the effect of λ on the PES. As a result, the energy barriers between different states are lowered and thus transitions between them are facilitated. This behavior is the analogous to sampling at high temperatures. Under sMD conditions, the canonical probability distribution for a given state of the system is modified to:

$$p^*(x) = e^{-\lambda V(x)/k_B T} \tag{18}$$

Where V(x) is the potential energy along a reaction coordinate x, k_B is the Boltzmann constant and T is the temperature.

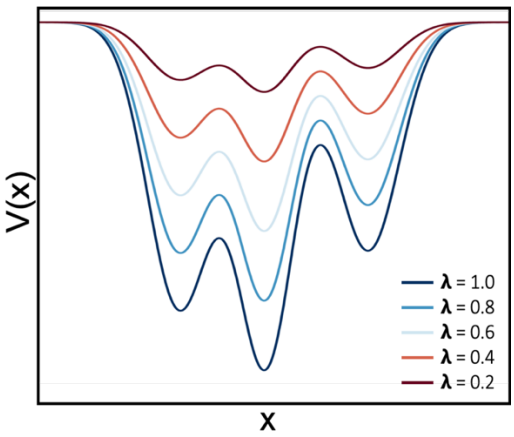


Figure 3. Effects of the scaling factor in the PES of the system. As more aggressive λ values are applied, the potential energy profile increasingly smoothed.

In this implementation from Sinko et al., a population-based reweighting scheme was proposed to reconstruct the canonical distribution of populations:

$$p(x) = p^*(x)^{1/\lambda} \tag{19}$$

In principle it is possible to consider other reweighting schemes, including for instance an energy term. However, it was shown that the population-based strategy is more accurate, as energetic terms are subjected to larger energy fluctuations that introduce large errors.[74]

2.3.2 Well-tempered metadynamics

MetaD is a well-known enhanced sampling method for studying rare events.[70] In MetaD the sampling is enhanced by adding a history-dependent bias potential constructed by depositing Gaussian kernels on selected degrees of freedom or CVs, that essentially describe slow degrees of freedom of the system. The bias potential added through MetaD is expressed by means of the following function:

$$V_G(s(x),t) = w\sum_{t'=\tau_G,2\tau_G}^{t} \exp\left(-\frac{\left(s(x)-s\left(x(t')\right)\right)^2}{2\sigma_s^2}\right) \tag{20}$$

where, w is the height of the Gaussian distributions, σ_s is the width of the Gaussian potentials, and τ_G is the time frequency at which Gaussian potentials are deposited along the CVs of a microscopic coordinate x of the system, s(x). These three parameters determine the accuracy and efficiency of

the free energy reconstruction. If the Gaussians are large, the FES will be explored at a fast pace, but the reconstructed profile will be affected by large errors. Instead, if the Gaussian are small or are deposited infrequently, the reconstruction will be more accurate, but it will take a longer time. The bias potential fills the minima in the FES, allowing the system to efficiently explore the space defined by the CVs. Figure 5 reports an example how MetaD acts with a one-dimensional bias potential.

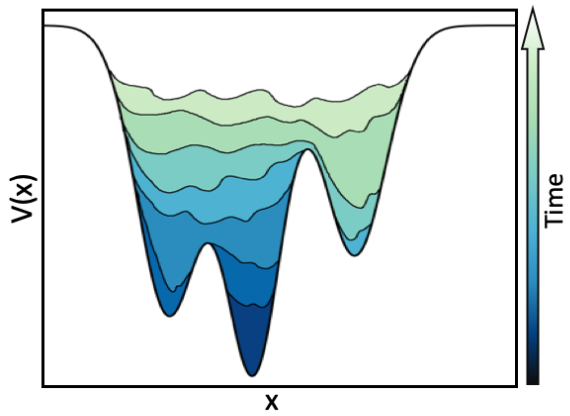


Figure 5. Pictorial representation of the MetaD method. The bias is gradually deposited at increasing time intervals τ along a x. Assuming that a simulation starts from the deepest basin, the gradual filling due to Gaussian deposition allows crossing the barriers and visiting the second deepest basin. The less profound basin is sampled after the other two basins are filled and the system is able to cross the barrier separating from them. Once all of the relevant minima have been visited, the system is free to sample along the entire reaction coordinate. At this point, minus the total bias accumulated gives the free energy.

In MetaD it is assumed that for a sufficiently long simulation, the system is able to freely diffuse in the CV space. One can reconstruct the free energy $F(s(x))$ in such CV space as:

$$\lim_{t\to\infty} V_G(s(x),t) = -F\big(s(x)\big) + C \tag{21}$$

Where C is an irrelevant additive constant. The main drawbacks of using MetaD are related to assessing the convergence of the simulation and choosing the CVs to bias.[76] As for the first one, once all the basins are visited, and while the simulation keeps running, the bias continues being deposited. This has the effect of overfilling the underlying FES and encouraging the system to visit high-energy regions of the CVs space. Thus, for a reliable FES estimate, the simulation should be stopped as soon as the system starts diffusing in the CVs space. A solution to this problem is provided by well-tempered MetaD.[77] While in standard MetaD Gaussians of constant height w are deposited over time, in well-tempered MetaD the bias deposition rate decreases over simulation

time. As a result, the overfilling mentioned above is less pronounced. Now, the Gaussian height becomes a function of the simulation time, according to:

$$w(t) = w_0 e^{-V_G(s(x),t)/k_B \Delta T} \tag{22}$$

Where w_0 is the initial Gaussian height, $V_G(s(x),t)$ is the total bias deposited at time t and ΔT is the range of the temperature at which the CVs are sampled. Each time the system is brought inside a new basin, the initial Gaussian height w_0 is restored and the simulation time-dependent scaling of the hills restarted. As the result, the bias potential tends to smoothly converge in the long time limit. Particularly important is the choice of the entity of decrease in Gaussian height per time unit. w should not become too small before a basin is completely filled, otherwise the system would remain stuck inside the basin with no possibility to overcame barriers. This can be controlled by setting for the simulation a specific parameter, the bias factor, defined as:

$$\gamma = \frac{T + \Delta T}{T} \tag{23}$$

Where ΔT is the upper limit of the temperature range to which the sampling of the CVs is confined. Thus, in the long-time limit, the total bias potential smoothly converges without fully compensating the underlying free energy as reported:

$$V_G\big(s(x)\big) = -\frac{\Delta T}{T+\Delta T} F\big(s(x)\big) + C \tag{24}$$

The second drawback that is encountered when MetaD is employed is the identification of an appropriate set of CVs. As already discussed, all of the slow degrees of freedom of the system should to be included by the CVs. Otherwise, the simulation will not converge, and the system will remain stuck at a certain position until the rare event involving the hidden CV eventually takes place. It is for this reason that choosing a proper set of CVs is far from trivial and knowledge about the system under study is required. In order to reduce the possibility of neglecting relevant degrees of freedom, several strategies can be applied. Among those relying on the use of CVs, one is using bias-exchange MetaD[78] in which multiple replicas of the system are simulated in parallel, and a different CV is biased in each replica. Another approach, specifically devised to manage particularly complex reaction pathways, is the use of path collective variable CVs (PCVs) .[79]

2.3.3 Steered MD (SMD)

In SMD simulations, a time-dependent external force is applied to a specific set of CVs in order to facilitate the sampling of the event of interest, which usually is achieved by MD simulation in not affordable computational time. In other words, SMD operates by pulling one or more CVs to guide the system from an initial configuration to a final one. In particular, an harmonic time dependent potential $U(x,t)$ acting on a CV *s(r)* is added to the standard Hamiltonian:

$$U(x,t) = \frac{k}{2}[s(x) - s_0(t) - vt]^2 \quad (25)$$

Where s_0 is the value of the CV at time 0, *t* is the time, and *k* is the strength of the external force applied. The value of the applied constraint force *F* is thus:

$$F(t) = -k[s(x) - s_0(t) - vt] \quad (26)$$

And finally, the external work ΔW performed on the system is derived from the *F(t)* by integrating the power along the time of the entire process:

$$\Delta W = v\int_{t_0}^{t_{final}} F(t)dt \quad (27)$$

Since ΔW is the cumulative change of the Hamiltonian in time and it related also to the change in energy of the system, it is potentially possible to calculate the free energy by guiding the system out of equilibrium between two states of interest by using methods based on the Jarkzynski equality.[80–82]

2.3.4 The MD-Binding Approach

As briefly introduced, the MD-Binding approach is here exploited to generate plausible protein-ligand binding trajectories at a reasonable computational cost. The MD-binding approach is a novel method, introduced by Spitaleri et al. in 2018[83] and available in BiKi Life Sciences,[84]. In this method an artificial electrostatic bias between the interacting partners (the ligand and a set of residues of the binding site) is introduced in order to guide the ligand to the binding site preventing the ligand to spend a lot of time in the bulk solvent. This idea comes from the well-known fact for which electrostatic interactions play a central role in the context of molecular recognition.[85]
The attractive external electrostatic bias acts between a subset of the residues of the binding site and the ligand with a functional form:

$$C \sum_{a \in A, b \in B} \frac{Q_a Q_b}{\tau_{a,b}} d(r_{a,b}, \lambda) \quad (28)$$

Where *A* and *B* are the two interacting partners, the ligand and the protein, and *a* and *b* are the atoms indices in *A* and *B*, respectively, $r_{a,b}$ is the distance between them. *C* is a modulation parameter, and $d(r_{a,b}, \lambda)$ is a function that modifies the effect of the bias according to the distance *r* and the parameter λ (see below). The charges, Q_a, Q_b, conventionally have opposite sign, are unitary and are spread among all the atoms of each set of atoms to let the system move as natural as possible without forcing unusual interactions.[83] The other key aspect of the MD-binding approach is the adaptivity of the bias. The adaptive bias lets the ligand to gradually approach, as gently as possible, to its naturally occurring binding mode avoiding unphysical pathways. The adaptive behavior of the external bias is reached by modulating the intensity of the external force to be always a fraction of the intensity of the physical forces naturally felt by the ligand. Moreover, the biasing force is gradually lowered as the process moves forward, which means that the ligand is moving toward its binding site. Finally, the biasing force cannot be turned back on, once switched completely off, even if the ligand leaves the binding site moving further to the solvent.
Mathematically, all these conditions are satisfied by modulating the coefficient *C* of Formula 28, along the simulation. In particular the modulus of the additive forces acting on the subset of ligand atoms is kept at a fraction of the modulus of the overall force created by the gradient of the regular potential energy of the system. In addition, $d(r_{a,b}, \lambda)$ of Formula 28 assumes the form of an exponential decay function according to:

$$d\left(r_{a,b}, \lambda\right) = \exp\left(\frac{-r}{\lambda}\right) \quad (29)$$

The switch-off is obtained via a scaling pre-factor γ that modulates the biasing force, calculated via a switching function as it follows:

$$\gamma = \frac{1}{1+\exp\left(-ss*(dist-th)\right)} \quad (30)$$

Where *ss* and *th* are two parameters and *dist* is evaluated as follows:

$$dist = min_{x \in A} min_{y \in \bar{B}} ds(x, y) \quad (31)$$

Where $\bar{B}$ is a subset of atoms of the selected residues of the binding site (*B*), known as switch-off atoms, $ds(x, y)$ is the pairwise distance between the two atoms *x* and *y*. According to this equation,

the bias is switched-off when any atom of the ligand A falls below a predefined distance from any atom of $\bar{B}$.

The advantage of using the MD-Binding approach to investigate protein-ligand event is the possibility of depicting the binding process as a whole, from the initial state in which the ligand is fully solvated in the bulk, to the final bound state without neglecting metastable states in reduced time compared with long unbiased MD simulations.

2.4 Path Collective Variables (PCVs)

As already discussed, nowadays it is possible to observe rare events such as the protein-ligand binding process at fully atomistic level leveraging enhanced sampling methods. One of the major drawback of enhanced sampling methods that rely on the use of the CVs is that a particular attention should be paid to the identification of these to properly guide the sampling.[76] Ideally the chosen set of CVs has to be representative of the process of interest and capable of describing it exhaustively without neglecting other fundamental variables that have to be considered in order to characterize the process in all its aspects. In principle, it is necessary to be deeply familiar with the process of interest in order to consider the best set of CVs, but this is the most of times infeasible. Indeed, every chosen CV has several requirements:

- It has to be a function of the coordinates of the system
- It has to be defined by a continuous mathematical function with continuous and analytical derivatives
- It needs to include all the slow modes of the system
- It has to be able of discriminating among initial, intermediate and final states

Moreover the number of employed CVs needs to be limited in order to save computational time.[67,70,86]

To overcame the difficulty of managing highly dimensional phase spaces and reduce the choice between multiple possible CVs, the PCVs formalism has been developed.[79] Consider a transition between state A and B, and suppose one is able to describe this reaction with a series of frames capturing the system at intermediate states along the reaction of interest. One can exploit this frameset to guide sampling along the "path" by means of the following CVs:

$$s(X) = \frac{1}{P-1}\left(\frac{\sum_{i=1}^{P}(i-1)e^{-\lambda(X-X(i))}}{\sum_{i=1}^{P}e^{-\lambda(X-X(i))}}\right) \quad (32)$$

$$z(X) = -\frac{1}{\lambda} ln\left(\sum_{i=1}^{P} \lambda e^{-\lambda(X-X(i))}\right) \quad (33)$$

For any microscopic configuration X the variable s can range between 0 and P, where P represents the number of frames comprised in the frameset. The difference $(X - X(i))$ is the distance between the configuration X of the system and the one adopted in frame i. It is common practice to use as distance metric the mean square deviation (MSD) and the CV, to be meaningful, requires the frames to be equidistant in the MSD space. Whenever the microscopic configuration sits on, that is corresponds to, a specific frame i, then all of the other terms in the summation disappear and $s(X) = i$. Thus, in practice, s describes the progression along the frameset. As for the second parameter, z can be though as orthogonal to s, and expresses the distance from the putative pathway. While one advances along the putative pathway, z allows exploring adjacent regions of the phase space. In the above functions, λ is a tunable parameter that ensures a continuous progression. It is proportional to the inverse of the average MSD between subsequent frames in the frameset. As a rule of thumb, the following formula it has been suggested:

$$\lambda = \frac{2.3(P-1)}{\sum_{i=1}^{P-1} |X(i)-X(i+1)|} \quad (34)$$

In summary, the highly dimensional space is reduced to a 1-dimensional description exploiting as CVs the progression along the pathway. The main advantage from combining s to z is a more permissive exploration of the configurational space around the guess route. As such, in the reconstructed FES, the minimum free energy pathway can be then recognized. There is no general rule to obtain the required frameset, but it has to respond to specific requisites. First of all, consecutive frames need to describe unidirectional progression towards the final state. No loops leading back and forth should be present. Secondly, equal spacing between frames is required. Finally, an appropriate number of frames should be chosen, so that the distance between subsequent frames is not excessive. Indeed, the resolution of the reconstructed FES is going to reflect the amplitude of the achieved spacing.

2.5 Analysis of multiple conformations: Cluster Analysis

Methods based on MD simulations have proven to be able to reproduce at atomistic level of detail the dynamics of complex biological system, such as protein folding and drug-receptor interaction. With the improvement of computer power, increasingly long trajectories consisting of large amount of data are collected. MD trajectories are sequences of data that specify the history of the atomic motions in terms of a sequential time-dependent set of configurations and the relative derived properties.[87] In the context of protein ligand binding, identifying those configurations that are

sampled by the system more frequently is of fundamental importance, since they (in good part) could be conformationally redundant. Data-mining techniques represent powerful tools to extract patterns and gain useful information from long MD trajectories. Among these techniques, cluster analysis[88] is one of the most widely used methods that allows to organize points of any data collection in different groups, called clusters, built according to a function measuring pairwise distances. The elements of a cluster are hence more similar to each other than those belong to other clusters. When clustering is applied to MD trajectories, the idea is to group similar configurations of the system sampled during the simulation. Since the sampling follows the Boltzmann distribution, where substates associated to low energy are more populated than those associated to high energy, clusters of configurations are differently populated. The parameters that have to be taken into account when clustering is applied to MD trajectories, are the choice of atoms subject to clustering, the pairwise metric used to compare elements, and finally the clustering algorithm. It is important to note that it is practically impossible to identify a perfect algorithm that produce a series of homogeneously clusters as the problem is NP-complete and thus unfeasible. Thus, in general, the quality of a clustering algorithm can be evaluated by its ability of identify the minimum number of clusters that preserve their internal homogeneity.

A wide number of different clustering methods has been developed and demonstrated to be very useful to identify the most visited states in MD trajectories. Among these, the most used clustering techniques can be classified as hierarchical and non-hierarchical.

2.5.1 Hierarchical algorithms

Hierarchical clustering algorithms can be agglomerative or divisive. In the first case they merge elements initially placed in different separate clusters, in successively larger clusters (bottom-up). In the second they begin with a single cluster encompassing the whole elements and proceed to divide it into successively smaller clusters (top-down). The results are reported in a descending or ascending “tree” diagram called dendrogram (see Figure 6), which is built as the algorithm iterates the process. The final output of a clustering algorithm can be derived from cutting the dendrogram at a chosen level and analyzing the corresponding number of clusters obtained.

These operations rely on the use of an appropriate metric (a measure of distance between pairs of observations), and a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets. In this work to analyze MD simulations we used the Root Mean Square Deviation (RMSD) as metric. Among the possible linkage criteria, the most commonly used are single-linkage, average-linkage, centroid-distance and complete-linkage. Single-linkage defines the distance between clusters as the shortest intercluster point-to-point distance; average-linkage, measures cluster-to-cluster distance as the average of all the distances between each point of the two clusters; centroid-linkage is similar to single linkage but the

intercluster distance is calculated between the cluster centroids; complete-linkage identifies cluster-to-cluster distance as the largest point-to-point intercluster distance between two clusters. A distance threshold is then defined, under which the conformations are considered as belonging to the cluster. In hierarchical algorithms, the number of clusters cannot be decided a priori, the clustering procedure can simply be stopped when a meaningful number of clusters has been reached.

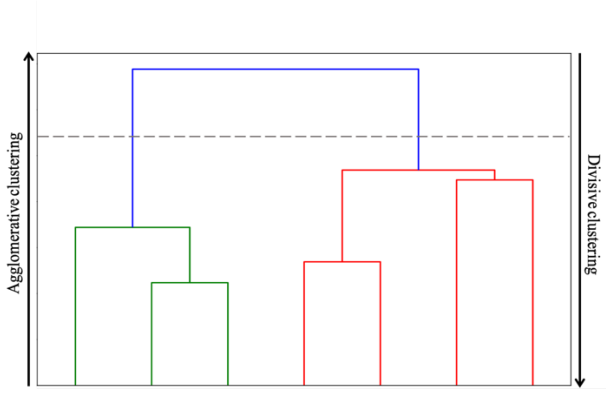


Figure 6. Hierarchical dendrogram. Clusters are represented in different color by cutting the dendrogram at the level of the grey line. The two rows on the left and right of the dendrogram schematize the agglomerative and divisive clustering approaches.

2.5.2 Non-hierarchical algorithms

Non-hierarchical algorithms define a classification by partitioning a dataset in a set of separate, non-overlapping clusters without hierarchical relationships among them. The most used algorithms to cluster MD trajectories are *k-means* and *k-medoids*.[89,90] *k-means* clustering partitions n objects in a number *k* defined *a priory* of clusters. The k centroids are determined in an iterative way; at each iteration every sample is assigned to the nearest centroid, which coordinates corresponds to the mean of the coordinates of the objects in the same cluster; this iteration is self-consistently iterated till convergence. As pairwise metric it is convenient to use the Euclidean distance of RMSD. Because the algorithm operates with means, which identifies coordinates that have minimal sum of squared deviations from it, *k-means* algorithm minimizes a distance-based cost function, which is the sum squared errors, SSE and is by defined:

$$SSE = \sum_{c=1}^{k} \sum_{i=1}^{n} |x_i^c - m_c|^2 \quad (35)$$

Where k is the number of clusters, n is the number of configurations, x is the i*th* element of the c*th* cluster with mean m. One of the main drawbacks of *k-means* algorithm is that is very sensitive to

outliers and to the clusters initialization. A variant of *k-means* is represented by the *k-medoids* algorithm. *k-medoids* differs from k-means in considering as centroids the most centrally located real object of each cluster.

2.6 Applications of MD-based techniques to the study of the protein-ligand binding process

2.6.1 Plain MD

The earliest attempt to simulate the spontaneous protein-ligand binding event using unbiased MD simulations was made by Shan et al. in 2011, where two Src kinase inhibitors were randomly placed together with their target protein in a simulation box and let them free to diffuse to their binding site.[52] In this work, aside from reproducing several times complexes nearly identical to those resolved by X-ray crystallography, the authors also identified an unknown allosteric pocket, highlighting the potential of MD simulations as promising tool in standard drug discovery programs. Likewise, Buch and co-workers used long plain MD simulations to completely reconstruct in terms of the pathway and related energetics, the benzamidine to trypsin binding process.[51] Other examples were conducted for membrane receptors (the β-adrenergic GPCRs coupled to agonist and antagonist small molecules[53] and spontaneous binding of tiotropium and acetylcholine to M2/M3 muscarinic receptors[91]) and in recent times, for the purine nucleoside phosphorylase enzyme and its inhibitor, the transition state analog DADMe-immucilin-H.[92] Furthermore, the same method applied to the context of fragment-to-lead development demonstrated the wide range of the unbiased MD simulations applicability.[93] In this case as well Authors were able to reproduce all the crystallographic poses of carboxy-tiophene fragments present in the X-ray structure of AmpC β-lactamase; distinct binding modes were discriminated from both thermodynamic and kinetic standpoint, through analysis of the simulation.[94] Particularly important is the work of Dror and co-workers, where they studied the M2 muscarinic acetylcholine receptor along with a number of experimentally identified allosteric modulators for which no crystal structure was available, demonstrating how dynamic docking could be useful for revealing binding sites and poses of binders without no a priori information.[95]
In this section are included also those methods that rely on discontinuous trajectories indicated as "discontinuous approaches" since the dynamics is not enhanced through the use of external forces beyond the normal force field.

Adaptive sampling methods are a class of unbiased MD approaches based on the Markov State Model (MSM) framework. [96,97] These methods aim at saving computational time by replacing few

long plain MD simulations with many short ones, which start from under-sampled states identified along the binding process. In first implementations, human supervision was necessary to manually select from where restarting a simulation, but in 2014, Doerr and De Fabritiis introduced a completely automated protocol for reconstructing the binding process for the trypsin-benzamidine complex, achieving a converged binding affinity one order of magnitude faster than classical sampling.[98] This strategy is also implemented in a freely available environment (High Throughput MD, HTMD[99], where other adaptive sampling algorithms can be considered) and used it to study the cooperative mechanism of recognition of ionic cofactors and substrate in the myo-inositol monophosphatase enzyme,[100] and the binding process of the lipid inhibitor ML056 to sphingosine-1-phosphate receptor.[101]

Totally different approach is the supervised molecular dynamics (SuMD) protocol introduced by Sabbadin and Moro.[102] In suMD protocol, simulations that seem not to reproduce the binding event are discarded on-the-fly through the use of a tabu-like algorithm. In practice, at defined time intervals of short simulations windows, the distance between center of masses of the binding site and the ligand is recorded and, fitted in a linear function. When the slope of the resulting line is negative, the distance between center of masses is getting shorter and the ligand is moving toward its binding site, and the simulation is let free to evolve. Conversely, when the slope is positive, the simulation is stopped, and a new run is restarted from the previous saved coordinates through velocity reassignment. Finally, when the distance between protein and ligand is lower a certain user-defined value, continuous plain MD simulations is restored. The method has been applied to a number of complexes of both globular and membrane proteins,[103,104] demonstrating to be able to reproduce X-ray structure in nanoseconds timescales. However, to correctly apply suMD, *a priori* knowledge of the binding pocket is required, thus excludes the possibility to discover new binding sites like in long plain MD simulations. Moreover, because of the discontinuous nature of the simulations, the binding path followed by the ligand the most of times is not close to the minimum free energy pathway.[25]

Another class of methods are multiscale approaches. Here the idea is to limit the full atomistic resolution MD simulation only to those configurations in which the ligand is close to the binding site of the protein. Recently Zeller et al. introduced a multiscale approach to dynamic docking that also allows the calculation of kinetics of the binding process using as a test case two inhibitors of the H1N1 neuraminidase.[105] In this work, Brownian Dynamics was applied when the distance between binding site and ligand was greater than a properly defined value. In this region, ligands and protein are treated as rigid entities that undergo translational and rotational diffusion in implicit continuum solvent, limiting the applicability only for those cases where the binding pathway does not involve large conformational changes.

2.6.2 Enhanced sampling methods relying on CVs

One of the earliest attempts to characterize the ligand process exploiting an enhanced sampling method was conducted by Gervasio et al. in 2005.[106] The authors used MetaD, in which a history-dependent bias is added on a set of CVs to the underlying dynamics in order to discourage the system from exploring previously visited regions of the CV space (for further detail see section of Methods), to successfully reproduce the binding mode and the associated experimental binding free energy of four protein ligand complexes (β-trypsin/benzamidine, β-trypsin/chlorobenzamidine, immunoglobulin McPC-603/phosphocoline, cyclin-dependent kinase 2 (CDCK)/staurosporine). Later, Provasi et al. successfully applied MetaD on the study of the binding process of the nonselective antagonist naloxone to the alkaloid binding pocket of a delta opioid receptor, succeeding in correctly estimating the association constant from the reconstructed free energy profile.[107] This was made possible by confining the unbound state of the ligand using a conically-shaped restraint in order to efficiently limit the configurational space accessible to the ligand in the bulk region. The same concept was exploited by Limongelli et al. in 2013, in the development of the so-called funnel MetaD, where the ligand confinement in the unbound state was achieved by a funnel-shaped restraint.[108] Despite the successful results in reproducing the mechanism of ligand binding, MetaD suffers from the common limits of other CV-based methods, one of this is the need to choose an optimal set of CVs.[86] This makes difficult to apply MetaD to simulate systems characterized by a high degree of complexity.

In 2007, Laio et al. developed a new technique that combines concepts of MetaD and replica exchange, the bias-exchange metadynamics (BS-MetaD), with the intention of overcoming the limits of the traditional MetaD.[78] In BS-MetaD multiple MetaD simulations are performed in parallel and exchanged at interval period of time and each replica is biased with a time-dependent potential acting on a different CV. This alleviates the problem of CVs selection. In 2009, Pietrucci et al. employed BS-MetaD methods to describe the binding process of a small peptide to the HIV-1 protease.[109] Despite the simulation required 2 μs to converge, the author were able to accurately compute the free energy associated with ligand binding and unbinding as a function of 7 CVs and derive the kinetics of the process using a discrete-states kinetic model.

Reconnaissance metadynamics[110] (RMD) provides a valid alternative to BS-MetaD since a larger set of CVs is considered. RMD is a machine-learning approach where the algorithm adjusts the bias potential applied using data obtained from short plain MD simulations. This procedure helps the user to select the *a priori* set of CVs, providing a way to efficiently explore unknown mechanisms. In 2012 RMD was applied by Soderhjelm to identify and score protein-ligand binding poses of the well-known trypsin/benzamidine system.[110]

A largely different approach was proposed by Spitaleri et al. in 2018.[83] Under the assumption that the driving force of a the protein-ligand binding process can be modeled by an electrostatic force due to ligand and binding site opposite charges, Authors developed the MD-binding approach by applying an adaptative, electrostatics bias between the binding partners. The method was applied on a set of protein-ligand binding complexes of pharmacological interest in order to reproduce the ligand binding process in relative short (20 ns) simulation time and multiple replicas.[83]

2.6.3 Enhanced sampling methods based on tempering

Replica-Exchange MD[111] (REMD) is one of the most widely used tempering methods to enhance conformational sampling and consists on running independently several replicas in parallel of the same system, at different temperatures. At regular intervals, exchanges between adjacent pairs of replicas are accepted according to a Metropolis acceptance criterion. However, for this to happen, a large number of replicas is required to ensure a significant overlap of potential energies sampled at adjacent temperatures. A valid alternative is provided by Hamiltonian REMD (H-REMD), in which the different replicas are simulated at the same temperature while the system's force field is modified, ensuring a greater probability of exchanges between replica by varying only part of the systems among replicas.[112] This method was adopted by Luitz et al. to reproduce the binding modes for three different protein-ligand system: human FKBP protein (FKBP-52) in complex with the high affinity compound FK506 and with the lower affinity ligand SB3, the peptide-binding domain of murine MHC class 1 molecule in complex with a viral antigen, through explicit solvent simulations.[113]

Another tempering method is represented by accelerated MD (aMD). aMD improves the configurational space sampling by locally adding a non-negative boost potential to the system's potential energy that are below a certain threshold value, while leaving the rest unaltered.[72] Kappel et al. recently adopted aMD to simulate ligand binding process to the M3 muscarinic receptor.[114] In particular, long-timescale aMD simulations (hundred-nanosecond timescale) are used to identify the metastable ligand-binding sites of three molecules in a significantly shorter time than plain MD: the full agonist acetylcholine (ACh), the partial agonist arecoline, and the antagonist tiotropium.

In Multicanonical MD (McMD) a random walk sampling through the energy space is made possible by a bias applied to the system.[73] In this method, higher energy states and lower energy states have an equal probability of being sampled because regions at different temperatures defined by the bias are simulated simultaneously. An advantage of the McMD method is that the canonical ensemble can be reconstructed by a reweighting procedure at standard temperature 300 K easily. In 2007,

Kamiya et al. adopted McMD to successfully dock the inhibitor tri-N-acetyl-d-glucosamine.[115] Later, Bekker et al. performed long McMD simulations to identify the correct binding mode of the inhibitor CS3 to cyclin-dependent kinase 2 and after, to predict the binding free energy by TI in accordance with the experimental data.[116]

Selective integrated-tempering-sampling MD (SITMD) method, which is based on the integrated tempering sampling method, modifies the energy distribution in a simulation system through the introduction of a sum-over-temperature non-Boltzmann distribution factor.[87] This method, that is often used for protein conformational sampling, offers the possibility to selectively enhance he sampling of a particular object in a system. Recently, SITMD was successfully tested to explore the binding process for a trypsin-benzamidine complex by Shao et al.[117] by enforcing random walking for the ligand in the solvent and on the protein surface to facilitate its search for the binding site.

3 ORIGINAL CONTRIBUTIONS

3.1 Fully Flexible Docking via Reaction-Coordinate-Independent Molecular Dynamics Simulations

3.1.1 Introduction

Predicting the geometry of protein-ligand binding complexes represents an essential issue for structure-based drug design in order to speed up the process of identification and development of new drug candidates.[118] Molecular docking has become the method of choice to address this task because it is highly efficient and relatively straightforward to use.[119,120] Despite these advantages, common docking programs typically suffer from fundamental issues and limitations, especially when protein flexibility plays a relevant role during the association process or when the contribution of water molecules is essential.[121,122] For practical reasons, mostly due to computational resources, most docking programs ignore target rearrangement upon binding and also hydration. At best, they account for it through approximations.[35] The use of scoring functions, used as surrogates of binding free energy estimators, is another well-known source of inaccuracies by introducing several approximations[123–125] or by adopting convenient phenomenological descriptions (empirical and knowledge-based scoring functions).[126] To reliably predict protein-ligand binding geometries and related thermodynamics quantities, it is necessary to capture the full conformational flexibility of both binding partners and explicit solvent models. In this respect, MD simulations can provide a valuable alternative to conventional docking calculations since a complete description of the entire process of binding is obtained at different degrees of accuracy.[127] Thanks to the rapid development of hardware and software it is now possible, in principle, to observe the entire drug-binding process at a fully atomistic description via plain MD.[46,51,54,68] However, reaching an adequate sampling of the configurational space is still to computationally demanding and not affordable in a hit-to-lead or lead optimization steps of drug discovery process.[46] Over the years, several research groups have addressed these limitations using enhanced sampling methods in order to accelerate the investigation of rare events on affordable time scales.[54,67] These methods usually fall into two general categories: those that rely on *collective variable* (CV) biasing[68,70,128], and those are based on *tempering*[71–74,129]; which seem to be more suitable since they act by heating all degrees of freedom of the system or by modifying the Hamiltonian.[46] Here, we propose a general framework (Figure 7) for dynamic docking based on the sMD method.[74]

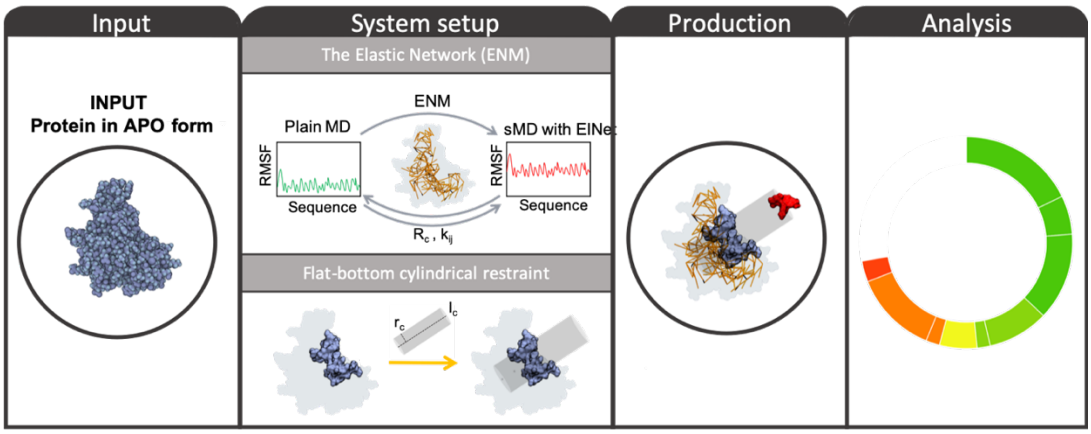


Figure 7. Schematic illustration of the dynamic docking workflow that consists of the following steps: 1) System setup, 2) Production and 3) Analysis. For further information see section 3.1.2 Methods.

As such, our approach is fully independent of any reaction coordinate or CV. It is unsupervised, fully automated, and can easily be implemented in several MD-based codes. The protocol was validated against its ability to reproduce the experimental binding modes of two pharmaceutically relevant systems:

- The glycogen synthase kinase 3β (GSK-3β),
- The N-terminal domain of heat shock protein 90-α (HSP90α), for which two different degrees of flexibility at the level of the binding site upon ligand binding exist. (Figure 8).

In particular, several studies show how the ATP-binding site of GSK-3β is adaptive, suggesting an open and closed state depending on the size of the ligand in order to increase the number of contacts.[130] On the contrary, the mechanism by which the N-terminal domain of HSP90α plasticity affects small molecule binding is not completely clear. It seems that the molecular recognition and binding of the residues composing the pocket with different small molecules involve both induced-fit and conformational selection mechanism. Indeed, the apo N-terminal domain of HSP90α may exist in different conformations, especially in a specific region at the entrance of the binding site where a particularly flexible segment is present, that undergo to internal motions upon ligand binding event modifying the population distribution towards more stable conformations.[131]

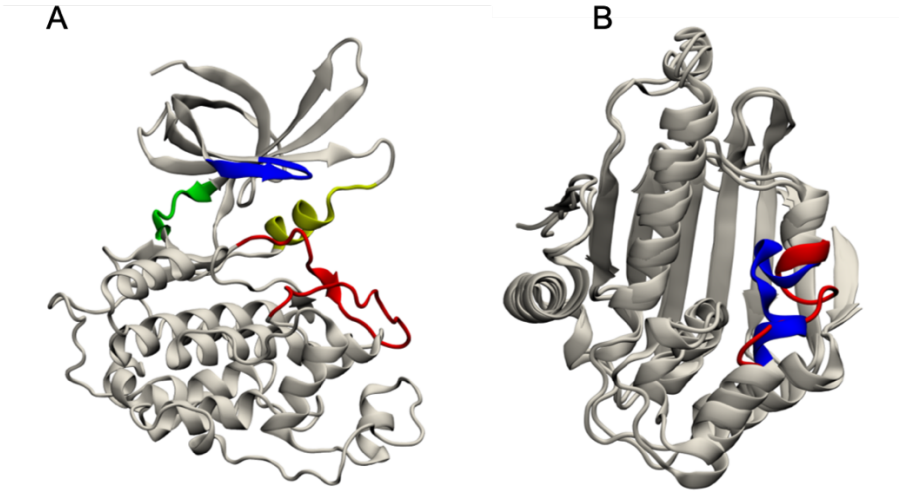


Figure 8. Crystal structures of the two test case. A) Apo form of GSK3β (PDB code: 1H8F). Highly conserved structural features, common to other kinases, that undergo conformational change through interactions with small molecules are depicted with different colors: helix αCin yellow, the activation segment is shown in red; the hinge region is represented in green and the glycine-rich loop is colored in blue. B) Both the crystal structures of apo-HSP90α are reported. The main difference is in the loop part of α-helix3 which adopts two different conformations here shown in different colors: "in" in blue and "out" in red (PDB code: 1YER and 1YES respectively).

GSK-3β is a proline-directed serine/threonine kinase involved in several physiological processes including glycogen metabolism and microtubule stability.[132] HSP90α is molecular chaperone with important roles in maintaining the functional stability of several client proteins.[133] Specifically, five ligands were considered for each system. Our procedure reproduced the experimental binding modes for all the cases examined. Moreover, is shown how the method can be used prospectively by exploiting a sufficient number of repeated simulations and taking advantage of appropriate data analysis.

3.1.2 Methods

Essentially, the overall protocol for dynamic docking here proposed and represented schematically in Figure 7, consists of the following three phases:

- **System setup**: starting from the apo structure of a protein, a 100 ns of plain MD simulations is performed in order to capture the internal motions of protein in aqueous environment under standard conditions when the ligand in not present. This information is required to implement the Elastic Network Model (ENM), whose function is preserving the whole

structure of the protein during the sMD simulations. At the same time, flat-bottomed cylindrical restraints are defined to limit the configurational space accessible to the unbound ligand.

- **Production**: a series of 10 sMD simulations with an intermediate scaling factor of $\lambda = 0.5$, lasting up to 100 ns each, are performed starting from the equilibrated apo structure of the protein and with the ligand placed 25 Å from the binding site in a random orientation.

- **Data Analysis**: identification of the most probable binding mode based on either a reference structure (retrospective analysis) or cluster analysis (prospective analysis).

3.1.2.1 System Setup

3.1.2.1.1 Identification of ligand-binding pockets

First of all, a fundamental prerequisite to correctly employ dynamic docking protocol is the identification of the pocket where the ligand is supposed to bind. To define the binding pocket, we take advantage of NanoShaper[134] a molecular surface tool available graphically in the BikiLifeScience Suite,[135] that applies a ray casting and grid-based method to find cavities and in general surfaces in a biological entity. In particular NanoShaper defines ligand-binding pockets on the initial PDB as the difference between two solvent excluded surface enclosed volumes generated with different probe radii. For HSP90 (PDB code: 1YER and 1YES),[136] the two radii used to compute the ATP-binding pocket were 1.4 Å and 4 Å, respectively, while for GSK3β (PDB code: 1H8F)[137] radii of 1.8 Å and 4 Å were employed.

3.1.2.1.2 Plain MD simulations setup

Preparation of each protein structure was performed through the "Protein Preparation Wizard" module of the Maestro interface (version 9.3.5) in Schrödinger suite.[138] The protein preparation process involved addition of hydrogen atoms, deletion of all crystal waters, and capping of protein termini. Plain MD were performed using version 4.6.5 of the GROMACS[139] software. The Amber ff99SB-ILDN force field[140] was employed to model the proteins. Solvation was achieved adding TIP3P[141] water molecules in a cubic box extending at least 12 Å beyond the protein surface. A proper number of water molecules was replaced with counterions for system neutralization (6 Cl^- for GSK3β, 7 Na^+ for HSP90α). Each system was energy minimized for 5000 steps of steepest

descent followed by gradual heating up to 300 K in an NVT equilibration, with steps of 100 K over 300 ps. Then, each system was equilibrated in the NPT ensemble at 300 K and 1 atm for 1 ns using a Parinello-Rahman barostat.[142] A cutoff of 10 Å and the PME method[62] were used for computing short-range interactions and long-range interactions, respectively. Constant temperature conditions were provided by using the v-rescale thermostat.[143] A time step of 2 fs was employed and all bonds to hydrogen atoms were constrained using the LINCS algorithm.[144] Production runs were carried out in the NVT ensemble at 300 K.

3.1.2.1.3 sMD simulations setup

Ligands were parameterized according to the general AMBER force field (GAFF)[145,146] for organic molecules using the Antechamber and parmchk tools implemented in Ambertools.[147] The Gaussian 03[148] package was engaged for geometrical optimization and calculation of the electrostatic potential of each ligand, applying the 6-31G* basis set at the Hartree-Fock level of theory. Partial charges were then calculated using the RESP method[149] as implemented in Antechamber. Using an in house script written in tcl for the Visual Molecular Dynamics (VMD) software,[150] each ligand was randomly placed at 25 Å from the binding pocket center of mass (COM) in line with the base of the cylindrical restraints defined using PLMUED 2.1 software[151] (see further). The same energy minimization and equilibration procedures reported for the plain MD setup were followed. The Amber ff99SB-ILDN force field was employed for the proteins, and sMD production was executed in the NVT ensemble at 300 K. All of the simulations were performed using an in-house version of GROMACS 4.6.7 modified to perform sMD.

3.1.2.1.3.1 Elastic Network Model (ENM)

Our ENM was implemented as a set of harmonic potentials of uniform and isotropic force constant k_{ij} acting between pairs of C_α atoms of the whole backbone, with exception of those residues that constitute the binding pocket, within a cut-off distance R_c.
The total harmonic potential of the ENM, $V_{ENM}(x)$ (where (x) represents the configurational space of the system) is given by:

$$V_{ENM}(x) = \frac{1}{2}\sum_{x_{ij}\leq R_c} k_{ij}(x_{ij} - x_{ij}^0)^2 \quad (36)$$

In equation 36, x_{ij} is the displacement between the i^{th} and the j^{th} C_α from their equilibrium distance x_{ij}^0 obtained from 100 ns of plain MD simulations as previously mentioned. The *g_distMat*[152] tool for GROMACS was used to compute the average equilibrium-distance matrix between C_α atoms from a 100 ns-long plain MD simulations. An in-house tcl script running under VMD was used to generate and to visualize the ENM basing on the C_α equilibrium-distance matrix and the parameter values (R_c, k_{ij}) specified by the user.

The resulting total potential energy function $V^*(x)$ scaled by sMD consists of two terms: the potential energy function proper of the force field $V(x)$, and the potential energy function related to ENM $V_{ENM}(x)$, as reported by the equation:

$$V^*(x) = \lambda(V(x) + V_{ENM}(x)) \tag{37}$$

The idea is to preserve the normal dynamics of the whole system and, at the same time, to enhance fluctuations of the residues composing the binding pocket in order to facilitate possible rearrangements of those residues that interact with the ligand upon binding. The ENM was validated through short sMD simulations of 10 ns. To this end, we compared the C_α-Root-Mean-Square Fluctuations (C_α-RMSF) obtained from trial sMD simulations and plain MD, which was acting as a reference. The ENM was considered satisfying when similar RMSF profiles could be obtained. Sets of parameters providing reasonable overlap were: $R_c = 1.0\,nm$ and $k_{ij} = 100\,kJ\,mol^{-1}nm^{-2}$ for both GSK3β and the N-terminal domain of HSP90α. A more thorough validation of the dynamics of the systems was performed by comparing the essential conformational sub-space. To quantify the difference, we computed the Root Mean Square Inner Product (RMSIP) between the first 10 eigenvectors obtained from each simulation as a measure of the overlap of the configurational explored by the two sets of simulations:

$$RMSIP = \frac{1}{D}\sum_{i=1}^{D}\sum_{j=1}^{D}(\eta_i{}^A\eta_j^B) \tag{38}$$

where η_i^A e η_j^B are the i^{th} and j^{th} eigenvectors obtained from Principal Component Analysis (PCA) to be compared, respectively, and *D* is the number of eigenvectors considered (in our case *D* = 10). RMSIP ranges from 0 to 1: the value is 1 if sampled subspaces are identical, and 0 if the sampled subspaces are completely orthogonal. In our simulations, the first 10 eigenvectors were sufficient to describe a significant fraction (more than 55%) of the total motion for both systems. The obtained RMSIP values of 0.74 for GSK3β and 0.70 for N-terminal domain of HSP90 demonstrated the high overlap between the essential subspace visited by the two sets of simulations.

3.1.2.1.3.2 Cylindrical Volumetric Confinement

The rationale of flat-bottomed restraints is to discourage the ligand from moving outside the volume of a cylinder centered on the binding site. The axis of the cylinder runs along the norm of the binding pocket, with one extremum extending within the pocket and the other one extending in the bulk, far enough to allow the ligand to be entirely solvated (25 Å from the COM of the binding pocket). The cylinder is constructed as follows. First, the norm of the binding pocket on the crystal structures is defined as the axis connecting the COM of the atoms composing the binding site and the COM of those atoms comprised in the entrance region of the pocket. NanoShaper was employed to unambiguously identify the two groups of atoms. Then, in order to avoid large fluctuations of the cylinder position during the sMD simulations, we defined the cylinder axis basing on the more stable portions of the protein as revealed from trial sMD simulations (C_α-RMSF < 1.5 Å). To this aim, we selected two groups among such atoms whose COMs lied along the previously defined norm (p_1 and p_2 in Figure 9). The whole procedure was carried out through an in-house python script.

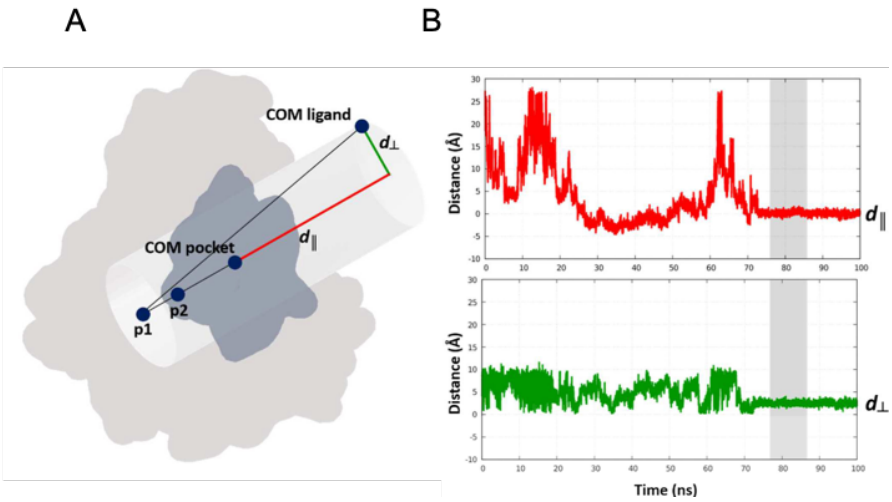


Figure 9. A) Representation of the cylindrical volumetric confinement. The norm of the binding pocket identifies the axis of the cylinder, which is defined as the line passing through two p_1 and p_2. p_1 and p_2 are the COMs, of two groups of Cα-atoms belonging to stable portions of the proteins. The length of the cylinder, $d_\parallel$, corresponds to the distance between the projection of the ligand COM on the axis of the cylinder and the COM of the binding pocket; while the radius of the cylinder, $d_\perp$, is the distance between ligand COM and its projection on the same axis. B) Definition of the metastable states from the time series of $d_\parallel$ and $d_\perp$. A metastable state is identified as any stable ligand configuration, described by its COM position with the respect to the cylinder, maintained for at least 10 ns. Example taken from run 7 of 2YI0, "in" conformational state of the protein.

The shape of the cylinder is defined by two flat-bottomed restraints with a force constant of 300 kJ mol^{-1} nm^{-2} each. In particular, the distance between the projection of the ligand COM on the axis and the point p_1 defines the length of the cylinder ($d_{\parallel}$ in Figure 9A), while the distance between the ligand COM and its projection on the axis define the radius of the cylinder ($d_{\perp}$ in Figure 9A). As previously mentioned, $d_{\parallel}$ was set to 25 Å from the COM of binding pocket. Conversely, the radius of the cylinder ($d_{\perp}$) was chosen as a trade-off between computational efficiency and the ability of not affecting the binding event. Specifically, on the one hand we wanted to avoid excluding any regions of the protein surrounding the binding site entrance that might interact with the ligand and contribute significantly to the binding process (larger values for $d_{\perp}$). On the other hand, we wanted to maintain the volume accessible to sampling as much limited as possible (smaller values for $d_{\perp}$). We tested different values for $d_{\perp}$ through trial simulations, which finally led us to determine an optimal radius of 8.1 Å for both systems. These flat-bottomed restraints were implemented exploiting the MATHEVAL facilities provided by the PLUMED 2.1 software. An example of the plumed file used is below.

Plumed file

```
WHOLEMOLECULES STRIDE=1 ENTITY0=protein ENTITY1=ligand

p1 : COM ATOMS=index of the CA-atoms protein whose center of mass is p1
p2 : COM ATOMS=index of the CA-atoms protein whose center of mass is p2
lig : COM ATOMS=index of heavy atoms of the ligand

alpha : ANGLE ATOMS=lig,p1,p2
d : DISTANCE ATOMS=lig,p1

MATHEVAL …
LABEL=d_parallel
ARG=alpha,d
VAR=a,d
FUNC=cos(a)*d
PERIODIC=NO
… MATHEVAL

MATHEVAL …
LABEL=d_perpendicular
ARG=alpha,d
VAR=a,d
FUNC=sin(a)*d
PERIODIC=NO
… MATHEVAL

#length of the cylinder
UPPER_WALLS ARG=d_parallel AT=length in nm KAPPA=300.0 LABEL=uwall1

#radius of the cylinder
UPPER_WALLS ARG=d_perpendicular AT=radius in nm KAPPA=300.0 LABEL=uwall2
```

3.1.2.1.4 Production

A series of 10 sMD production runs, lasting up to 100 ns each, were performed for each of the 5 protein-ligand complexes in order to ensure that the ligand binding event took place. For HSP90, 2 different conformational states of the protein (loop-in, loop-out) were considered. This resulted in a total number of 150 simulations (50 for loop-in HSP90, 50 for loop-out HSP90, and 50 for GSK3β).

3.1.2.1.5 Data Analysis

3.1.2.1.5.1 Retrospective Analysis

In order to assess whether ligand binding took place within the production runs of 100 ns of sMD simulations, we monitored the RMSD of ligand heavy atoms with respect to the crystal structure after optimal alignment on C_α protein atoms. The crystal binding pose was considered reproduced when a RMSD of 2.8 Å or less was observed and maintained for at least 10 ns (Figure 10). A metastable state was considered reached when a RMSD greater than 2.8 Å was maintained for at least 10 ns.

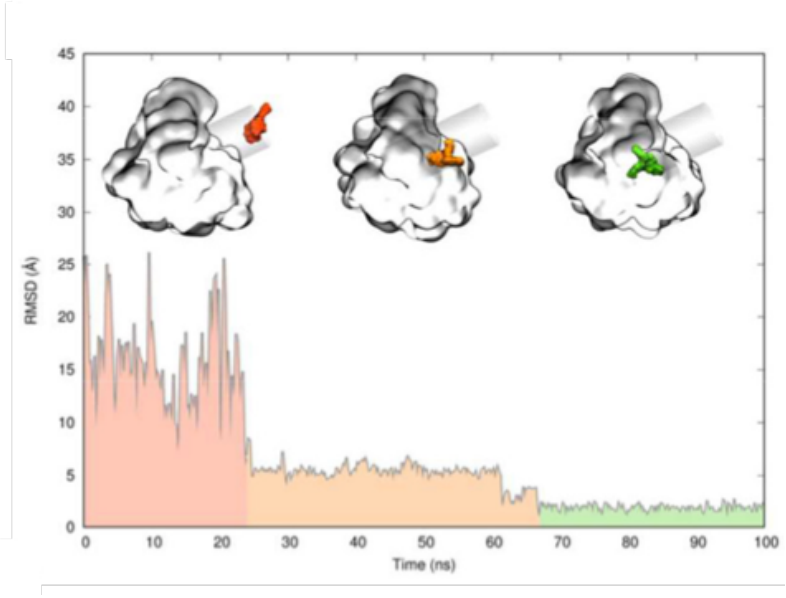


Figure 10. Time evolution of a ligand's RMSD from the corresponding crystal structure in a typical productive run. The experimental binding mode is considered reproduced if a RMSD lower than 2.8 Å was observed and maintained for at least 10 ns (green profile). A metastable state is taken in account when a RMSD greater than 2.8 Å was preserved for at least 10 ns (orange profile). No productive run is characterized by a not stable RMSD. Example taken from run5 of 2YI5, "out" conformational state of the protein.

3.1.2.1.5.2 Prospective Analysis

The previous approach can be exploited only when the crystal structure of the bound complex is available. Thus, we devised a procedure to identify relevant binding poses even in the absence of experimental data. First, we filtered the trajectory of each sMD production run in order to consider only relevant metastable states, that is relatively stable ligand configurations preserved for at least 10 ns. From a geometric standpoint, a metastable state was defined in terms of the position of ligand COM with respect to the axis of the cylinder ($d_{\parallel}$ and $d_{\perp}$). The ligand position was considered stable when the average values of $d_{\parallel}$ and $d_{\perp}$ varied at most of 0.5 Å during 10 ns of simulation, and 90% of the total fluctuations fell below 2.0 Å (Figure 9B, right panel).
For each protein-ligand system, the 10 ns of each run were joined together in a concatenated trajectory, on which the RMSD matrix was computed by least square fitting on the protein backbone C_{α} atoms and stored. The RMSD matrix was then processed by means of a hierarchical clustering algorithm,[87] using the average-linkage method as implemented in R version 3.3.2.[153] The hierarchical clustering algorithm (bottom-up) combines existing clusters, creating a hierarchical structure that reflects the order in which the clusters are merged. For our clustering procedure, we employed a cut-off value of 2.0 Å. To discard those conformations that were not frequently sampled during the simulations, we discarded those clusters which population fell below 2% of the total number of structures retained for the analysis. The representative of the most populated cluster represents the ligand binding pose predicted.

3.1.3 Results and discussion

The dynamic docking protocol here proposed, is basically based on performing repeated sMD simulations of protein-ligand complexes started from the unbound states and applying a tunable ENM to preserve the overall protein-folding and a cylindrical volumetric confinement in order to improve the exploration of the configurational space in the regions that are more relevant for the binding process saving computational resources. In sMD simulations, the sampling is enhanced by modifying the potential energy function of the force field by a scaling factor λ between 0, in which all the interactions are switched off, and 1, which corresponds to plain MD. By properly choosing λ, it is possible to efficiently cross energy barriers, thus accelerating the transitions between different configurational states of the system. As the result, simulating under scaled potential energy conditions means that all the degrees of freedom of the system are enhanced in a nonspecific way causing protein unfolding. Here, an intermediate scaling factor equal to 0.5 is used in order to adopt a more generalizable strategy, resulting as the best trade-off between accuracy and the possibility of observing binding events in the 100 ns of simulation. As reported in Figure 11 using

a scaling factor of λ=0.5, large fluctuations in the protein structure were observed, which were not detected for the same system in plain MD simulations. To preserve the overall protein-folding during sMD simulations an isotropic ENM was introduced. This ENM acted on all the Cα atoms except those belonging to the binding site which was kept fully flexible (see paragraph 3.1.2.1.3.1 of Methods for further details). As already mentioned in Methods, the ENM was parametrized through a reasonable iterate procedure of optimization so as to satisfy the average Cα pair distances measured by hundreds of nanoseconds of plain MD simulations of the protein in the unbound state.

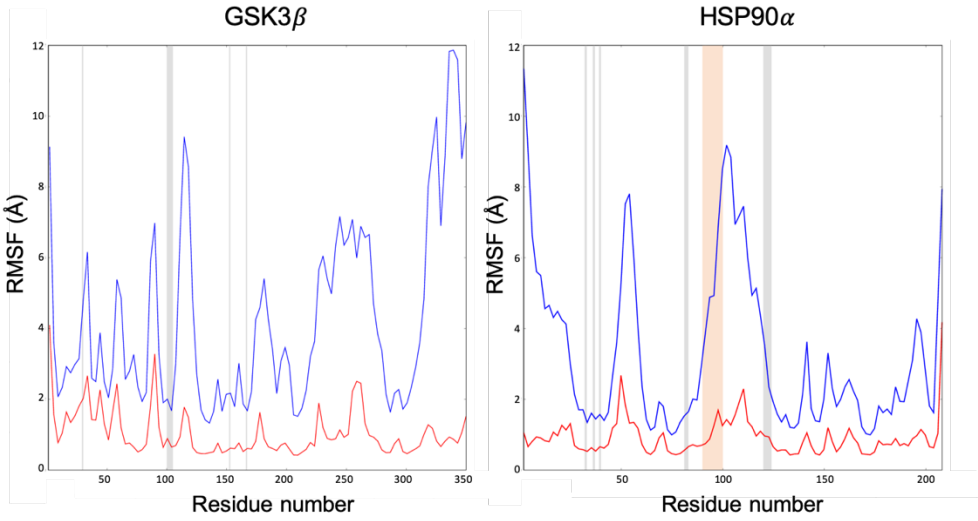
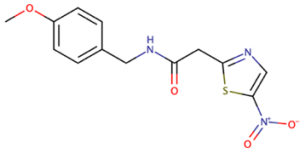
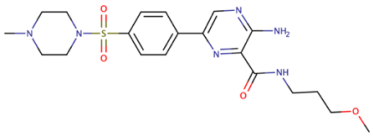
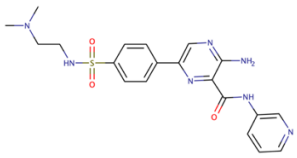
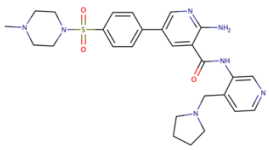
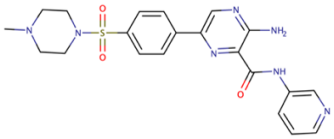


Figure 11. Comparison of the Cα-RMSF profile calculated between plain MD simulations (red) and sMD simulations with a scaling factor of 0.5 (blue). Residues forming the binding pocket are represented with the gray color for both the protein structures. Residues composing the loop in HSP90α structure are shown in orange.

As first test case, as already mentioned in the introduction, we employed GSK3β, a well-known pharmaceutical target. This proline-directed serine/threonine kinase plays a major role in several physiological processes, including glycogen metabolism and microtubule stability.[132] From a pathological standpoint, GSK3β is mostly involved in neurodegenerative conditions, such as Alzheimer's disease.[154] GSK3β is a two-domain enzyme: a N-terminal lobe (N-lobe), mostly consisting of β-sheets, and a large C-terminal lobe (C-lobe), essentially formed of α-helices. The ATP binding pocket is buried deeply within the two lobes.[130,155] We focused on five different reference crystal structures of human GSK3β in complex with selective inhibitors characterized by different receptor binding affinities. In particular, we selected four crystal structures of GSK3β in complex with pyrazine-derived inhibitors[156] (PDB code: 4ACD, 4ACC, 4ACG, 4ACD) and one structure in complex with a thiazole-derived inhibitor[157] (PDB code: 1Q5K) (Table 1). As already briefly introduced, the ATP-binding site of GSk3β, as the consequence of ligand interaction, may

undergo to different conformational changes of the most flexible features common to other kinases that are: the activation segment, the hinge region, the glycine-rich loop, and the helix αC^{130} (Figure 8A)

Table 1. Structures and biological activities of the five known inhibitors for GSK3β.

PDB	Structure	Biological Activity: K_i (nM)
1Q5K		38
4ACH		20
4ACC		20
4ACG		4.9
4ACD		0.22

In order to correctly set the parameters for ENM, we performed a series of short sMD simulations at fixed scaling factor of 0.5 with ENM imposed, using different combinations of cut-off distance R_c , and force constant k_{ij} (in Figure 12 is shown how different ENM models applied to GSK3β can adjust the dynamics when sMD simulation is performed).

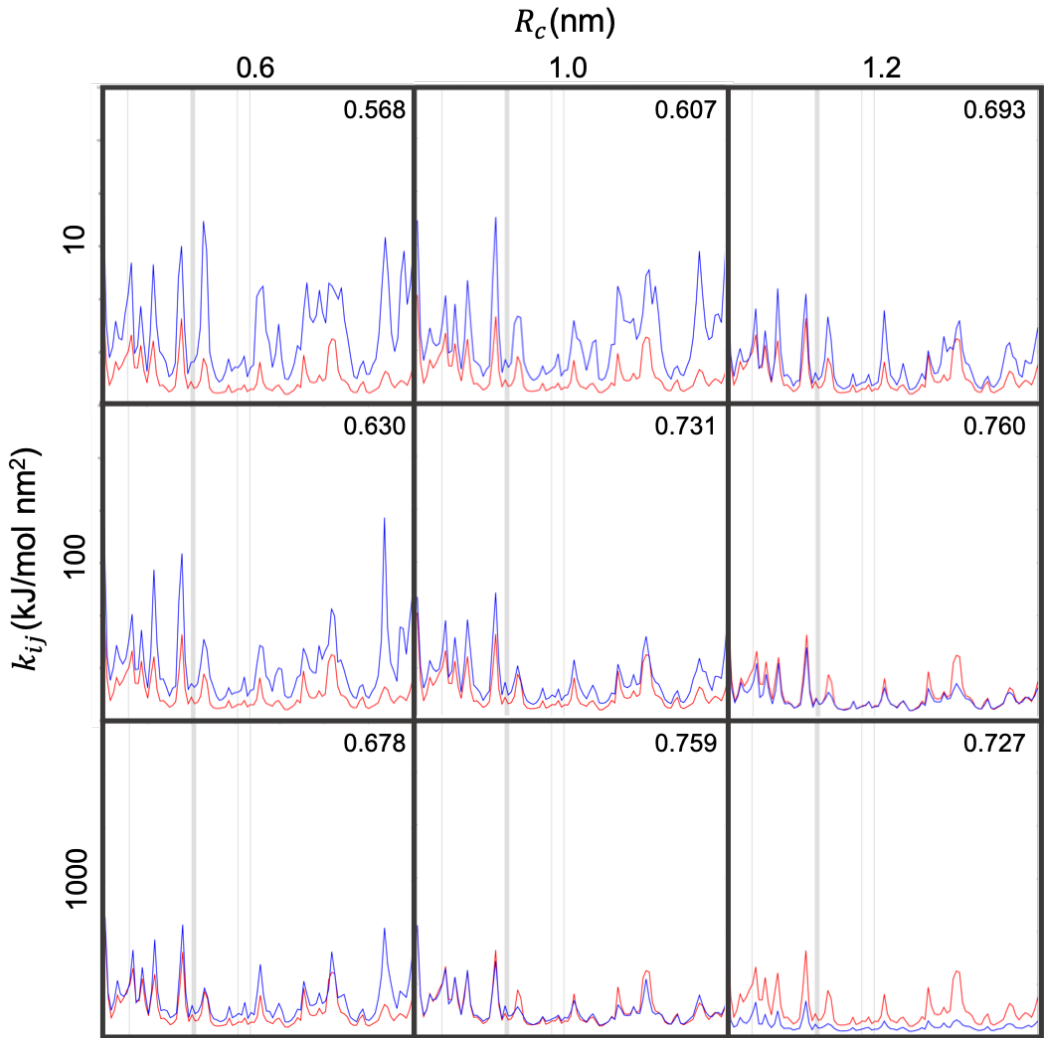


Figure 12. Effect of k_{ij} and R_c values on the structure of apo-GSK3β protein. Comparison of RMSF profile, of the Cα-atoms of the backbone as a function of residue number during 100 ns of plain MD (red profile) and 10 ns of sMD with ENM (blue profile). The RMSIP value for each set of parameters, computed between the first 10 eigenvectors obtained from plain MD and sMD simulations, is reported in in the top of each graph.

Set of parameters that provide a reasonable overlap, in terms of the RMSF, of the whole protein structure with the exception of the residues composing the active site was: $R_c = 1.0\ nm$ and $k_{ij} = 100\ kJ\ mol^{-1}nm^{-2}$.

In Figure 13, the outcome of the perspective analysis is reported for each run of all the individual protein-ligand complexes.

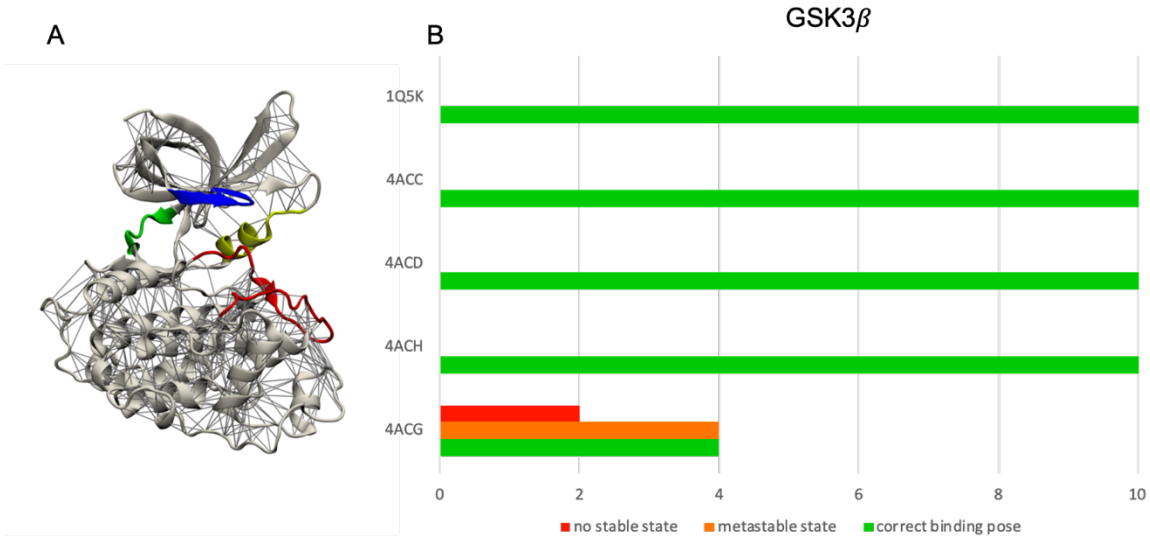


Figure 13. A) Pictorial representation of the ENM applied to GSK3β. B) Retrospective analysis of the 10 independent runs performed on GSK3β. Productive runs are represented in green, while runs where no stable state was reached within 100 ns of simulations refer to red bars. Runs ending in metastable states other than the experimental binding mode are reported in orange.

As indicated, in four out of the five studied complexes a 100% success rate was achieved. In other words, this means that in these cases, all of the 10 sMD runs correctly reproduced the experimental binding mode of the crystallography structure within 100 ns of sampling. In the cases of 4ACG, the success rate was reduced to 40%.

In order to characterize the metastable states sampled along the different trajectories, we performed cluster analysis of the aggregate trajectory (Figure 14). As reported, the center of the most populated cluster for 4ACG (cluster 1, population 25.7%) was only 2.64 Å away (in terms of RMSD) from the reference crystal structure and, surprisingly, all the pivot interactions with the protein were well reproduced. Indeed, the relatively high RMSD was mostly imputed to a different conformation adopted by the solvent expose piperazine moiety of the molecule (Figure 14) with no specific key interaction with molecular target.

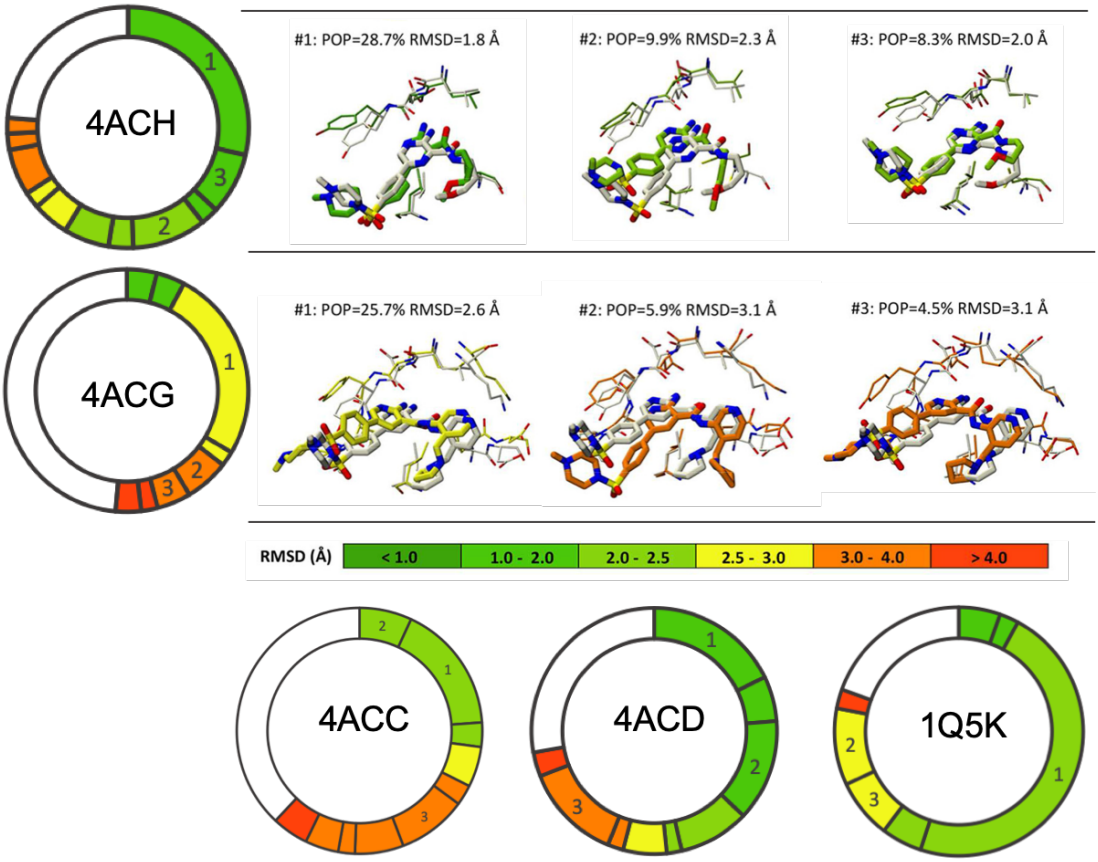


Figure 14. Validation of dynamic docking protocol for its ability to reproduce the experimental binding modes based on cluster analysis. Results of the cluster analysis of the metastable states for all the investigated complex for GSK3β are described in the pie charts, with color based on the RMSD between the representative structure of the cluster with respect to the corresponding crystal structure. For the best and the worst performing complexes (PDB code: 4ACH and 4ACG, respectively) the representative configurations from the top three populated clusters are represented. Each representative configuration is shown superimposed to the crystal structure (white) and color-coded according to its RMSD.

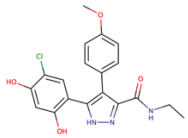
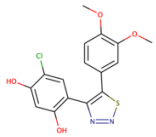
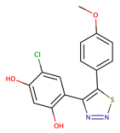
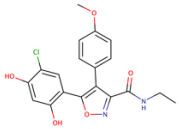
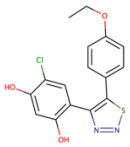
As the second test case, we considered the N-terminal domain of HSP90α, a molecular chaperone that takes part in many different processes with the aim of maintaining the functional stability of several client proteins.[133] HSP90α is also involved in many complex pathologies such as cancer, Alzheimer's and Parkinson's[158] diseases among others,[159] highlighting its relevance as a pharmaceutical target.[160] The N-terminal domain of HSP90α contains an ATP-binding site to which most of the known inhibitors bind, and is thus the most studied portion of the protein.[133]

As already introduced, there is a considerable flexibility observed in this site, which is characterized by an antiparallel β-sheet at the bottom and by three helices as its walls, one of which, the α3, is located at the entrance of the binding site. In several studies, this last helix has been observed to adopt distinct and ligand-specific conformations, including partly misfolded states in the middle

portion (residues 103 to 107). In particular, two main conformations are adopted by this region in the apo protein ("in" and "out" states, see Figure 8 B).[136,161]

Because crystal structures of these two limiting conformations are available in the PDB (codes: 1YES, 1YER)[136] we decided to consider both in order to apprehend if the ligand may be more selective to one specific conformation adopted by the protein than the other. We focused on five different reference crystal structures of human N-terminal domain of HSP90α in complex with selective inhibitors (Table 2). In particular, we selected three crystal structures in complex with thiadiazole-derived inhibitors (PDB codes: 2YI7, 2YI5, 2YI0),[162] one with a isoxazole-derive inhibitor (PDB code: 2UWD),[163] and one with a pyrazole-derived inhibitor (PDB code: 2BSM).[164]

Table 2. Structures and biological activities of the five known inhibitors for HSP90α.

PDB	Structure	Biological Activity: K_d (nM)
2BSM		78
2YI5		39
2YI0		7.5
2UWD		5
2YI7		4.8

The intrinsic dynamics of the apo N-terminal domain of HSP90α for each residue was calculated as the average of 100 ns of plain MD on both conformations available (1YES and 1YER). We follow the same procedure as the previous case to test different values of ENM parameters as show in Figure 15.

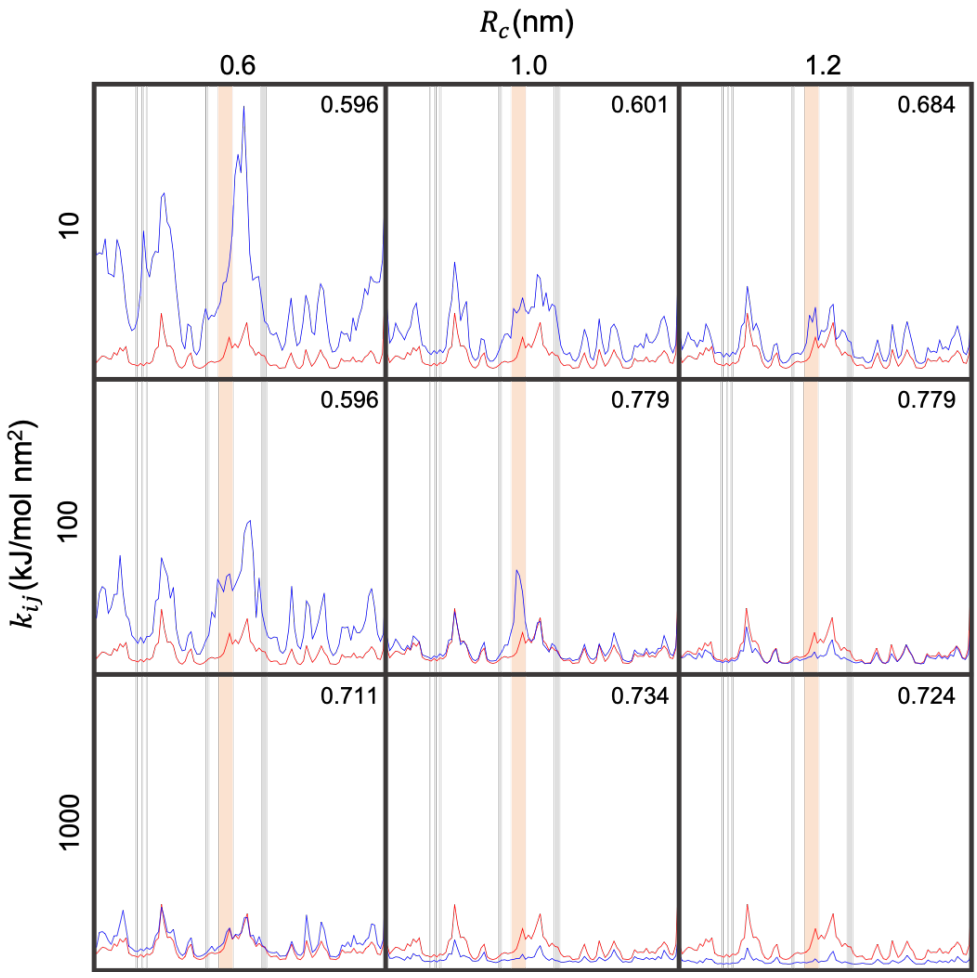


Figure 15. Effect of k_{ij} and R_c values on the structure of apo-HSP90α protein. Comparison of the RMSF profile, of the Cα-atoms of the backbone as a function of residue number during 100 ns of plain MD (red profile) and 10 ns of sMD with ENM (blue profile). Residues forming the binding pocket are represented in gray, while residues forming the loop at the entrance of the binding pocket are colored in orange. The RMSIP value for each set of parameters, computed between the first 10 eigenvectors obtained from plain MD and sMD simulations, is reported in in the top of each graph.

The Set of parameters that provide a reasonable overlap, in terms of the RMSF, of the whole protein structure with the exception of the residues composing the active site was: $R_c = 1.0\ nm$ and $k_{ij} = 100\ kJ\ mol^{-1}nm^{-2}$. In Figure 16 are reported the results of the retrospective analysis for the five ligands starting from apo form of HSP90α with both the "in" and "out" states. In this case, the results are more complicated to understand than the previous test case.

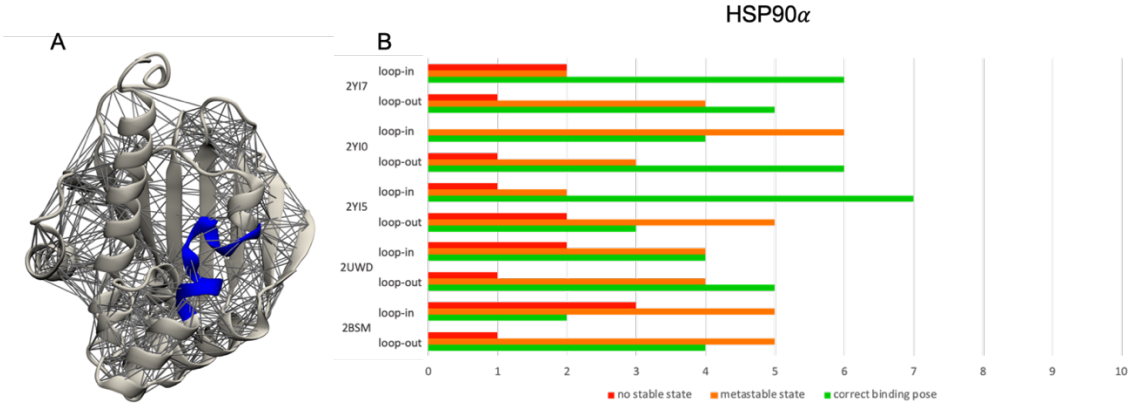


Figure 16. A) Pictorial representation of the ENM applied to the N-terminal domain of HSP90α. B) Retrospective analysis of the 10 independent runs performed on GSK3β. Productive runs are represented in green, while runs where no stable state was reached within 100 ns of simulations refer to red bars. Runs ending in metastable states other than the experimental binding mode are reported in orange. Example taken from "in" conformational state of the protein.

In particular, none of the individual complexes achieved a 100% success rate but, in the best scenario, only the 55% success rate was showed (2YI7). The less satisfying success rate (30%) was achieved in the worst situation (2BSM). This less satisfactory performances most likely reflected a more complex problem related to the inherent flexibility of the binding site. In spite of this, the binding mode of the crystal structure was replicated at least twice in all the studied complexes. Furthermore, the conformation of the helix $\alpha 3$ in the related crystal complex of reference was visited at last once for each system, with the exception of 2BSM, demonstrating that the procedure is robust enough to provide rather reliable results independently from the initial configuration of the protein. Although this suboptimal performance might weaken the reliability of our framework when employed prospectively, the obtained results show that this is not the case. Indeed, stable metastable states could be hardly reached by most of the unsuccessful runs. In other words, on average, "wrong binding modes" turned out to be transient states and, therefore, they affected only marginally the correct identification of the experimental geometry among the most populated clusters. This behavior is nicely captured by Figure 17, where we show the results of the metastable states cluster analysis for the best and worst performing investigated complexes (2YI7 and 2BSM, repsectively).

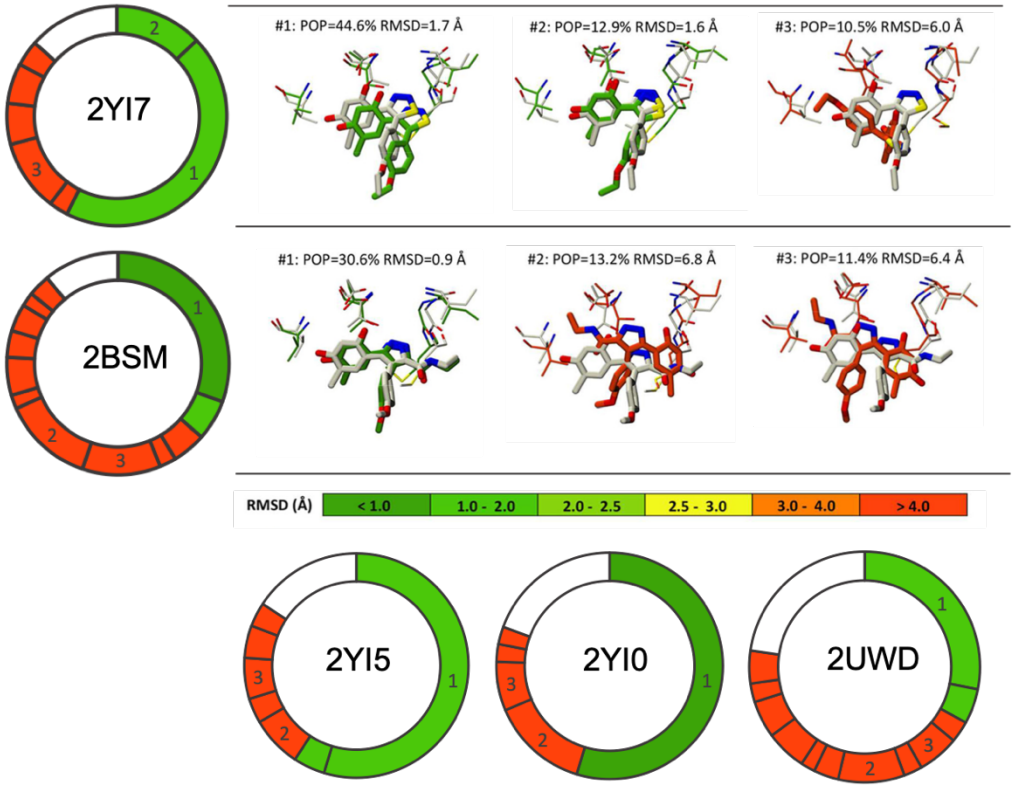


Figure 17. Validation of dynamic docking protocol for its ability to reproduce the experimental binding modes based on cluster analysis. Results of the cluster analysis of the metastable states for all the investigated complex for HSP90α are described in the pie charts, with color based on the RMSD between the representative structure of the cluster with respect to the corresponding crystal structure. For the best and the worst performing complexes (PDB codes: 2YI7 and 2BSM, respectively) the representative configurations from the top three populated clusters are represented. Each representative configuration is shown superimposed to the crystal structure (white) and color-coded according to its RMSD.

In the former, the most populated cluster (population: 44.6%) shows a binding mode with a RMSD lower than 1.7 Å from the experimental structure. Despite the lower success rate, in the case of 2BSM, the binding mode identified by our protocol was still found among the top ranked clusters (RMSD of 0.91 Å, populations 30.6%). Most importantly, this cluster was unambiguously separated in terms of relative population from the best ranked "wrong binding mode" (population 13.2%). Hence, we can safely conclude that, as long as the correct geometry is reproduced at least once in repeated sMD runs, our framework has a very good potential to successfully identify it among the most populated clusters

3.1.4 Conclusions

Summarizing, we presented an efficient framework to predict protein-ligand binding geometries through MD-based simulations. Remarkably, the overall protocol is based on commonly affordable time-scales, which are accessible to the vast majority of current research groups. Key ingredients of our approach are a CV-free enhanced sampling method, such as sMD, a tunable ENM and a cylindrical volumetric confinement. We showed that the proposed protocol is able to successfully identify the experimental binding modes among the top ranked clusters for all of the investigated cases. Being conceptually simple, moderately flexible, and of straightforward implementation in several MD-based platforms, we believe that this approach has a great potential in structure-based drug discovery, where a balanced tradeoff between accuracy and speed is essential.

3.2 Predicting residence time and drug unbinding pathway through sMD

3.2.1 Introduction

As already introduced in the Introduction section of this thesis, residence time, which corresponds to the period of time a drug stays in contact with its biological target, is potentially considered a key parameter at the early stages of drug discovery as important as affinity since is better correlates with in vivo efficacy.[2] Evaluating on- and off- rates at an experimental level requires expensive setup and time-consuming testing of new molecules in complex kinetic assays.[165] Concerning the prediction of binding kinetics through in-silico approaches, MD simulation is the method of choice to study the protein-ligand binding or unbinding process at a fully atomistic level.[25,166] However, simulating rare events such protein-ligand binding reaching an ideal trade-off between accuracy and efficiency is still far from a trivial task. In this respect, to accelerate the sampling of rare events in order to obtain an adequate statistics, multiscale simulations[167–169] and enhanced sampling methods have been introduced.

Although enhanced sampling methods are mainly used to obtain a thermodynamic characterization of the process, nowadays interest has been raised about recovering kinetic information from these methods. Among enhanced sampling methods which introduce an external biasing force acting on selected degrees of freedom, MetaD has been successfully applied to the trypsin-benzamidine system,[170] to describe the mechanistic unbinding process and the relative kinetics of Dasatinib from c-Src kinase[171] and to predict residence times for a series of congeneric compounds from the p38 MAP kinase.[172] Recently, Gobbo et al. introduced a novel approach based on adiabatic bias MD with an electrostatics-like collective variable to correctly rank a ligand series of glucokinase. Subsequently, the same approach was applied in a prospectively way to successfully predict residence time for a congeneric series of GSK3β inhibitors.[173] Among enhanced sampling methods based on tempering, τRAMD technique, in which forces are randomly assigned to the ligand promoting its expulsion from the binding pocket without specifying any CV,[174] was used by Kokh et al. to rank different compounds according to their dissociation time.[175] Moreover, sMD has been shown to be efficient in ranking congeneric series of ligands according to their residence times.[4,165,176] Recently, sMD has also been exploited to predict, without any a priori experimental information, the unbinding kinetics for a series of hDAAO inhibitors by Bernetti et al.[177]

In the current research described in the following, sMD was used to characterize the kinetic behavior of seven drug-like compounds, inhibitors of HSP90α. In particular, first a ranking of these ligands based on computed residence times is provided, then relevant features of the unbinding process are extracted through dimensionality reduction techniques and cluster analysis of the trajectories.

3.2.2 Methods

3.2.2.1 sMD simulation

All the protein-ligand system were setup using the BiKi software, version 1.3.5.[135] Every inhibitor molecule was geometrically optimized via a quantum mechanical approach computed at HF/6-31G* level of theory with NewChem.[178] Partial charges were derived using the RESP[149] method as implemented in Antechamber. The Amber99SBildn[140] and GAFF[145,146] force fields were used to treat respectively the protein and ligands. When the crystal structure of the protein-ligand complex was not available, the scaffold of the ligand was located in the binding pocket in agreement with the scaffold of its congeneric molecule for which the experimental structure exists. The protein-ligand complexes were placed in the geometrical center of a parallelepiped-shaped box of 7.58 nm length and all the boxes were solvated with TIP3P[141] water molecules. In order to preserve the electro-neutrality of the system, some water molecules were replaced with sodium ions. All the initial coordinates of each system were minimized in a 5000 step minimization run, using the steepest descendent method. A classical equilibration procedure of four steps was applied. In particular, the first three steps were run for 100 ps in the NVT ensemble, starting at a temperature of 100 K, increasing during the following steps up to 300 K. Positional restraints were applied to the heavy atoms of the protein backbone using an isotropic force constant of 1000 kJ/mol nm^2.
The fourth equilibration step was carried out for 1 ns in the NPT ensemble at 300 K and 1 atm. An integration step of 1 fs was employed in the first and in the final equilibration steps, while the remaining steps were performed using an integration step of 2 fs. The pressure-coupling was carried out using the Parrinello-Rahman barostat.[142] All simulations were performed using a customized Gromacs 4.6.1[179–181] version that was able to perform sMD, using the velocity Verlet algorithm[182] for the integration of equations of motion. The neighbor list cutoff within the Verlet cutoff scheme was set to 1.1 nm. PBC were applied and PME method[62] was used to treat Coulombic interactions with a grid size of 0.16 nm. The cutoff for long-range Coulomb interactions was set to 1.1 nm and short-range van der Waals interactions were treated with the same range cutoff of 1.1 nm. The Lincs algorithm[144] was used to restraint all the bonds of protein and ligands while the SEATTLE algorithm[183] was used for water molecules.
A series of 25 partially unrestrained repeated sMD production runs were performed for each complex, for a total of 175 simulations, until the occurrence of the unbinding event. A Python script running with VMD was employed to uniquely define the unbinding event and stop the simulations. In particular, the ligand was considered unbound when the distance between the protein COM and ligand COM achieved 30 Å.[150] sMD simulations were performed using a scaling factor λ of 0.45. In order to preserve the overall protein fold during the simulations, harmonic positional restraints were applied to the heavy atoms of the backbone. The residues composing the binding site, located

within 5 Å from the bound ligand, were not restrained at all in order to allow rearrangements in this region upon ligand unbinding.

The average unbinding times, corresponding to the computational residence times τ^{calc}, along with standard errors, standard deviations and bootstrapped standard errors, were calculated from the exit binding times for each ligand. We followed the procedure as described by Mollica et al.[165] to correlate the normalized computational residence times with the normalized experimental residence times τ^{exp}. With the bootstrapping method, several virtual samples set (unbinding times) of the equal size to the original one, are generated and the mean is calculated. By these averages it is possible to estimate the standard deviation of the mean itself. R studio was used to perform the Bootstrapping procedure.[184]

3.2.2.2 Postprocessing of unbinding trajectories and measure of similarity between different pathways

In order to assess similarities and difference between different unbinding pathways, a clean-up stage of the trajectories followed by the projection of these into a low-D dimensional space is required. Notably, the low-D dimensional space is capable of well-describing relevant features of the unbinding event that occur in the high-dimensional configurational space. During the cleanup stage of the data analysis, the trajectories are postprocessed in order to store only those frames whose RMSD of all heavy atoms of the ligand after alignment on the protein's alpha carbons, is equal or greater to a given threshold compared to the previous frames stored. Specifically:

- If the RMSD calculated between the current frame of a specific trajectory ($j_m; m = 1, 2, 3, \ldots, M$) and all the N previously saved frames ($i_n; n = 1, 2, 3, \ldots, N$) is greater or equal to the given threshold δ, then the current frame is saved
- If the RMSD between the current frame j_m and one frame belonging to the stored trajectories i_n is lower than δ, the i_n is replaced by the running frame j_m and all the subsequent structures of the stored trajectory are discarded (Figure 18 A).

In this work, we considered an RMSD threshold value of 2.0 Å for compounds 1 and 7 and 2.5 Å for the remaining compounds. The choice of using different values for δ reflects the evidence that compounds with a greater amount of degrees of freedom (compound 2 and 5) require an higher RMSD threshold to better describe their unbinding pathway. The final results of the cleanup stage is that only productive transitions are preserved while common useless features such as loops and/or dead ends along the unbinding pathway are filtered out (Figure 18 B). Once the trajectory is postprocessed, the low-D dimensional space was constructed through the Isomap algorithm[185], a

nonlinear dimensionality reduction technique, as implemented in Pysomap python library (Figure 18 C). Since Isomap attempts to preserve the geodesic distance of the high dimensional space (high-D space), it represents a superior approach, compared to other multidimensional scaling methods MDS (both metric and nonmetric) which consider only the Euclidean distance of the high-D space, because both the pairwise distances and the intrinsic geometry of the high-D space are preserved. The Isomap algorithm consists of three-steps. Considering the manifold M of the high-D space, the first step defines which points are neighbors on M based on the distances $d_x(i,j)$ between pairs of points i,j in the input space X.

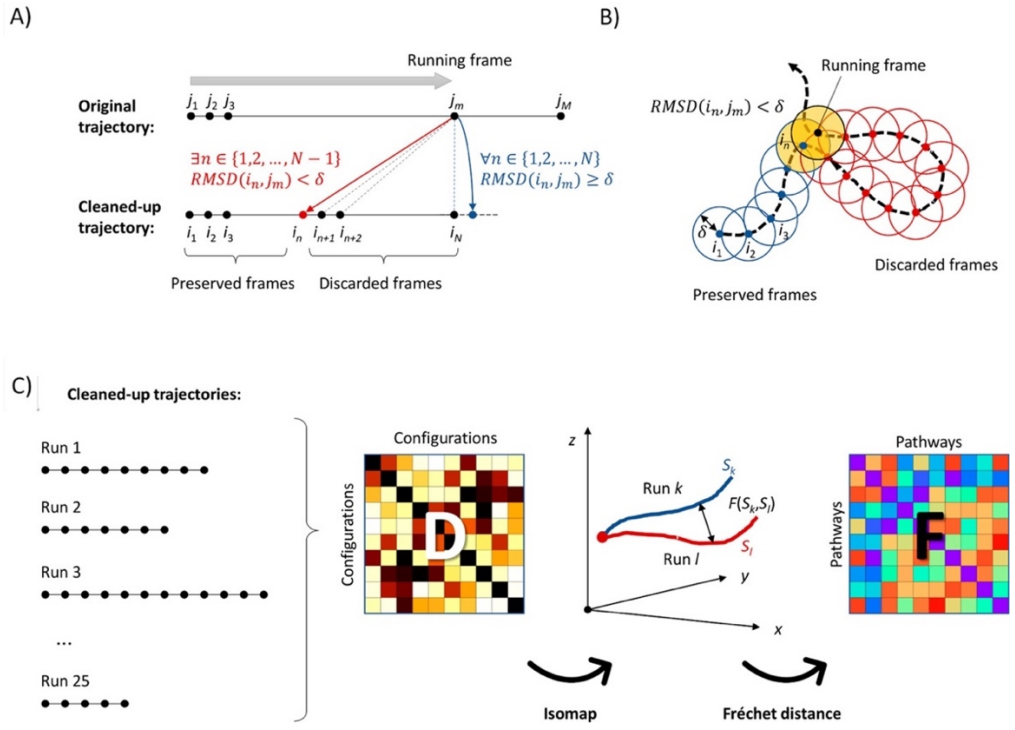


Figure 18. A) Schematic representation of the cleanup stage of trajectories. B) Illustration of a loop feature in a hypothetical trajectory that is removed during the cleanup stage. C) After cleanup stage, the pairwise distance obtained in the configurational space of the ligand is computed using all the frames of each preserved trajectory (matrix D in the illustration) and embedded in the low-D space through Isomap. Then, each unbinding pathway is projected un the low-D space. Finally, for each pair of unbinding pathways, the Fréchet distance is computed and stored in a symmetric matrix (F in the scheme) that can be further exploited for subsequent cluster analysis.

A weighted graph G over the data points, where the weight of the edges is $d_x(i,j)$, is defined. In the second step, in order to determine the geodesic distances $d_M(i,j)$ between all pairs of points on M, the shortest path distances $d_G(i,j)$ in the graph are calculated. In the final step the classical

MDS algorithm is applied to the matrix of graph distances to get a d-dimensional Euclidean space. Here, we used the distance between the ligand's heavy atom coordinates computed over all pairs of trajectory frames as a metric.[185] We decided to embed each postprocessed trajectory into a 3D space using the 10 k-nearest neighbor method; this strategy was defined via a trial and error procedure and a comparison of the features of the unbinding pathways projected in the low-D space. Even though a high density of points in the high-D space is a fundamental prerequisite for the ability of Isomap to correctly reproduce the geodesic distances, it becomes very difficult to handle with Isomap big data such as MD trajectories. Thus we used a landmark variant of Isomap that helps to reduce the entire data set, considering only a fraction of the entire unbinding trajectories to build the initial embedding.[186] In this specific case, we used a landmark-like approach, in which only the states in the high-D spaced describing a gradual advancement from the bound to the final state in which ligands are completely solvated were considered. This procedure turns out to be particularly advantageous because the embedding is built only on the sub-manifold that better describes in terms of mechanistic information the unbinding events. Moreover, since sMD is equivalent to incrementing the temperature, the trajectories obtained through sMD are extremely noisy. As such, this procedure could be used to filter out all those features in the high-D space that corrupt the continuous advancement along the considered path such as: detours, dead-ends, loops etc.[187]

Once the unbinding pathways were projected into the low-D space built by Pysomap, the pairwise similarity between different paths was measured using the Fréchet distance, as implemented in the Similarity Measures library of R.[188] The discrete Fréchet distance is a measure of similarity between curves that considers the position and the order of the points along two curves of different lengths. In particular, the discrete Fréchet distance between two paths, S_k and S_l, is defined as the infimum over all parametrization α and β of the maximum distance t of the same curves as reported by

$$F(S_k, S_l) = \inf_{\alpha,b} \max_{t\in[0,1]} [d(S_k(\alpha(t)), S_l(\beta(t)))] \quad (39)$$

Where d is the Euclidean distance between frames of the parametric curves $S_k(\alpha)$ and $S_l(\beta)$ estimated in the low-D space. Using the Fréchet distance as dissimilarity metric, we were able to apply cluster analysis to group similar unbinding trajectories. Here we employed two cluster algorithms as implemented in R version 3.3.2 depending on the purpose of the analysis.[184] In particular, an agglomerative hierarchical clustering algorithm with the complete linkage method was used to group similar unbinding trajectories in *k* different clusters. The representative pathway of each cluster is represented by the one whose distance with all the other pathways in the cluster is minimum.

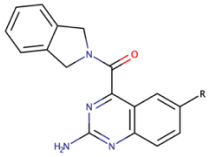
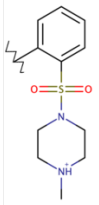
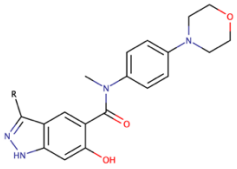
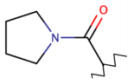
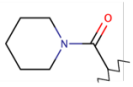
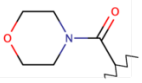
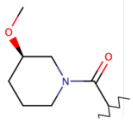
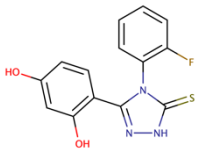
3.2.3 Results and Discussion

3.2.3.1 Test Case: HSP90α

As for a previous research, we selected the N-terminal domain of HSP90α as a test case in complex with seven different inhibitors since experimental SPR measurements were available. Moreover, for five of these inhibitors the crystal structure of the complex was available (compounds 2, 4, 6 and 7). As described in Methods section, the bound states of the remaining compounds (compounds 1 and 5) was modeled directly in silico, because of their minor structural transformations with respect to existing X-ray complexes. All the small molecules considered in this work are reported in Table 3 with the corresponding experimental binding affinity (equilibrium dissociation constant, K_d) and kinetic (association and dissociation rate constants, k_{on} and k_{off} respectively, and the experimental residence time τ^{exp}) observables.

As shown in Table 3, compounds 1 and 2 exhibit a quinazoline scaffold, compounds 3, 4, 5 and 6 have an indazole-like structure while compound 7 is a pyrazole-containing, resorcinol-like inhibitor. Despite the considered molecules possess very different scaffolds, all these molecules bind to the ATP-binding site of the N-terminal domain of HSP90α in a similar manner. To simulate the unbinding process through sMD simulations, we considered the initial state as the one with the ligand bound to its molecular target. 25 sMD independent runs for each protein-ligand complex were performed using a scaling factor of 0.45 in order to observe the unbinding event within few hundreds of nanoseconds for all the studied compounds. In order to preserve the overall structure of the protein during the sMD simulations, we applied weak positional restraints to all the backbone heavy atoms with the exception for those that form the binding pocket as described in Methods.

Table 3. Structures of the compounds investigated

compound	scaffold	substitution	K_d (nM)	k_{on} ($M^{-1}s^{-1}$)	k_{off} (s^{-1})	τ^{exp} (s)
1		H	2152.00	$2.56 \cdot 10^{5}$	$5.52 \cdot 10^{-1}$	1.81
2			13.00	$5.28 \cdot 10^{4}$	$6.80 \cdot 10^{-4}$	1470.59
3			18.73	$1.30 \cdot 10^{5}$	$2.43 \cdot 10^{-3}$	412.20
4			42.00	$2.14 \cdot 10^{4}$	$9.10 \cdot 10^{-4}$	1098.90
5			485.00	$2.72 \cdot 10^{4}$	$1.32 \cdot 10^{-2}$	75.76
6			195.30	$1.65 \cdot 10^{3}$	$2.94 \cdot 10^{-4}$	3402.52
7			98.00	$1.13 \cdot 10^{6}$	$1.11 \cdot 10^{-1}$	9.01

3.2.3.2 Ligand ranking based on the computed residence times

The computational residence time (τ^{calc}) for each compound was calculated as the average of the exit times observed for each ligand in all the independent 25 runs as reported in Table 4.

Table 4. Simulated exit times (in ns) for 25 independent sMD runs per compound.

	compound						
run	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	2.41	630.88	39.54	30.82	66.85	42.52	15.36
2	14.42	42.51	498.30	219.15	155.57	10.76	51.11
3	9.90	11.35	28.05	202.74	46.55	272.98	48.10
4	8.18	65.43	214.79	159.59	30.89	64.87	197.61
5	4.83	593.35	222.52	258.10	443.51	183.97	50.52
6	3.68	395.19	56.02	109.90	8.37	33.59	2.84
7	9.03	89.87	348.91	25.54	80.12	154.69	83.37
8	4.20	680.39	355.95	234.69	36.33	321.98	60.90
9	3.73	238.33	518.74	236.35	251.85	164.20	56.94
10	48.14	121.58	44.92	302.92	230.63	215.66	90.57
11	10.55	122.81	100.20	323.71	250.29	245.73	43.46
12	1.87	182.66	92.98	167.72	210.86	30.09	29.91
13	3.01	77.89	22.00	75.37	244.84	95.16	20.13
14	6.19	21.16	12.45	61.28	14.19	127.50	127.44
15	0.87	357.65	350.19	91.99	14.63	31.68	108.11
16	8.67	290.36	87.33	98.24	430.86	146.02	135.91
17	19.97	54.65	140.17	104.71	81.47	499.04	122.37
18	7.93	164.76	304.36	212.15	226.66	173.06	146.95
19	15.72	764.19	163.69	325.87	46.09	92.45	56.88
20	2.41	931.93	81.15	327.71	26.20	288.59	80.39
21	3.17	253.82	106.96	159.14	143.63	557.94	80.72
22	1.12	606.88	39.78	380.76	165.34	169.36	58.13
23	2.56	299.98	231.65	445.89	312.10	43.59	76.72
24	7.46	227.88	167.67	289.37	9.75	216.01	65.90
25	1.60	52.22	15.85	399.29	6.82	105.68	121.19
Average	**8.06**	**291.11**	**169.81**	**209.72**	**141.38**	**171.48**	**77.30**

As it is reported in Table 5, where all the computed residence times are compared with the related experimental residence times (τ^{calc} and τ^{exp}, respectively) for each of the seven compounds, we were able to estimate the exit times for all the compounds, with the exception of compound 6, in overall agreement with experimental data.

Table 5. Experimental and calculated residence times (τ^{exp} and τ^{calc}) for each compound.

		Computed quantities			
compound	$\boldsymbol{\tau^{exp} \pm \sigma}$	$\boldsymbol{\tau^{calc}}$	$\boldsymbol{\pm\sigma}$	$\boldsymbol{\pm\sigma e}$	$\boldsymbol{\pm\sigma_{BS}}$
1	1.81 ± 0.18	8.06	± 9.68	±1.94	± 1.77
2	1470.59	291.11	± 259.04	± 52.88	± 49.96
3	412.2 ± 13.43	169.81	± 150.50	± 30.72	± 29.20
4	1098.9 ± 320.64	209.72	± 115.53	± 23.58	± 22.53
5	75.76 ± 3.91	141.38	± 133.01	± 27.15	± 25.86
6	3402.25 ± 177.06	171.48	± 138.89	± 28.35	± 26.61
7	9.01	77.30	± 44.88	± 9.16	± 8.69

Indeed, the structure of compound 6 is the most different with respect to the other three compounds that present the indazole-like structure as scaffold (compounds 3,4,5). In particular, compound 6 shares a piperidine moiety with compound 4, with in addition a methoxy group. The addition of a polar oxygen of the methoxy group in this position, makes this region very similar, in terms of polarity, to the compound 5 that possesses the morpholine ring. From computational results, it seems that the addition of the methoxy group accelerates the dissociation of the ligand from the protein, similarly as for compound 5, in contrast to experimental data. Because all these substituents of this series (compounds 3-6) tend to occupy similar regions of the binding site, the bulkier substituent groups are capable to make additional interactions with the hydrophobic part of the pocket. In particular, according to experimental results, the methoxy substituent of the compound 6 adapts the methyl group to the hydrophobic pocket resulting in a more stable complex with respect to the compound 5, whose polar oxygen in the morpholine ring is more exposed to the solvent and consequently unfavorable to interact with the hydrophobic pocket, and compound 4. In our simulation, compound 6 was predicted as an in-between condition of compounds 4 and 5. However, this disagreement could be due to limitations of the force-field in discriminating these small details that might have been intensified by the choice of an aggressive scaling factor.

Furthermore, evaluating the degree of correlation between simulations and experiments for all the entire set of compounds (compounds 1-7), it appears clear from the plot in Figure 19 that the estimation of the computational residence time for compound 6 affected the correlation. In particular, we performed a linear regression procedure in which the effect of the scaling factor was considered by normalizing both experimental and computational residence times according to: $\left(\frac{\tau_i^{exp}}{\tau_{max}^{exp}}\right)^{\lambda}$ and $\left(\frac{\tau_i^{calc}}{\tau_{max}^{calc}}\right)$ respectively.

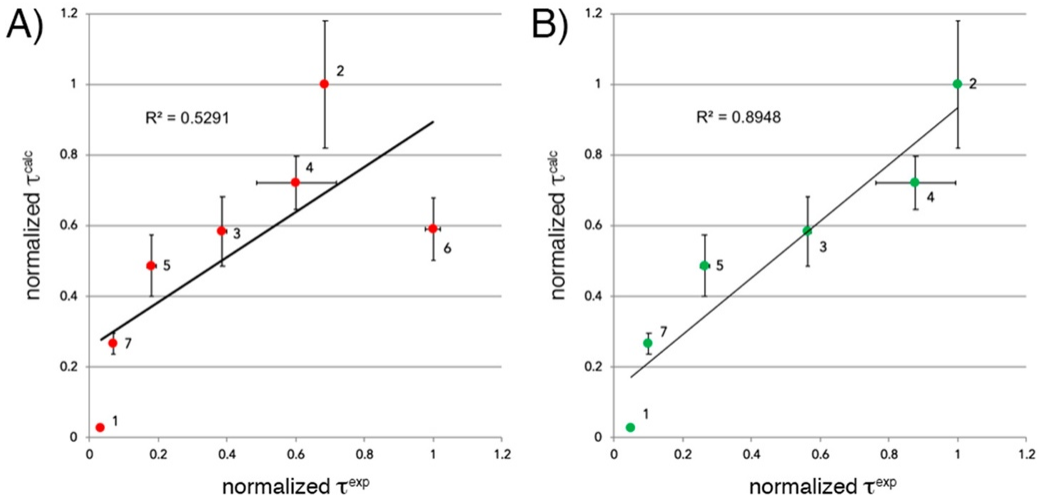


Figure 19. Normalized computed residence times plotted against normalized experimental residence times. Error bars correspond to the standard deviations reported in Table 5, σ. A) Results obtained considering the entire data set of compounds. Pearson correlation coefficients of R = 0.73 and R^2 = 0.53 were obtained. B) Results obtained excluding the compound 6, considered an outlier. Now, the ranking for the remaining 6 compounds was reproduced correctly. Spearman rank correlation of 1 with a Pearson correlation coefficient R^2 = 0.89.

Considering the entire set of compounds resulted in a Spearman ρ of 0.89, while p-values were significant ($P < 0.05$). Pearson's product-moment correlation resulted in an R of 0.73 and R^2 of 0.53 (Figure 19 A). Exploration of the studentized residuals, resulting in a value of $t > 3$ identified compound 6 as an outlier. The same linear regression procedure was applied to the same set of compounds excluding it (Figure 19 B). Now, the Spearman rank correlation gave a ρ of 1, while p-values remained significant ($P < 0.01$). Now, Pearson's product-moment correlation resulted in an R of 0.94 and R^2 = 0.89.

3.2.3.3 Similarity analysis of unbinding pathways

As introduced before, characterizing the possible pathways that a small molecule can follow to bind a specific target binding site and, in this way, to modulate it biological activity is of primary importance for drug discovery. However, multiple and different routes can be involved in unbinding processes, mostly due to different shape and size of the binding pocket and the residues faced along the path, the degrees of freedom of both the ligand and the protein. Here we tried to recognize and classify different routes from sMD simulations in order to recover, if possible, significant mechanistic details about the unbinding event and describe structure-kinetics

correlations. As described in Methods, we embed in a 3D space the Cartesian coordinates of the ligand's heavy atoms along each trajectory, taking advantage of the dimensionality reduction technique Isomap. The similarity between different trajectories was assessed in the low-D space calculating the discrete Fréchet distance, which is the metric employed in the following cluster analysis. Determining the similarities between different trajectories in the low-D space is relatively straightforward compared to analyze each trajectory in the high-D space. Notably, through this analysis of similarity in the low-D space based on the Fréchet distance of all the 25 sMD runs for each ligand, we were able to quickly identify different unbinding routes. In Figure 20 is shown an example of the strategy devised, using the trajectories of compounds 1.

All the significant configurations visited by the compound 1 in all the 25 different unbinding trajectories were used as an input for the dimensionality reduction procedure (clean-up stage and Isomap). Thus, each trajectory could be followed in the new low-D space identified by the projection of all the configurations in this space. Finally, all the Fréchet distances evaluated between all of the trajectories are stored in a similarity matrix on which a hierarchical clustering was performed in order to identify similar pathways. In this specific case, according to Figure 20 we could identify, for instance, two pairs of paths that resulted to be very different in the low-D space (couple 1: path 3 and 16, couple 2: path14 and 18). In particular, from the dendrogram exhibiting the hierarchical relationships between trajectories, three major clusters could be identified cutting the dendrogram at height of 9. Paths 4 and 18 belong to the same largest cluster, while paths 3 and 16 are part of the second cluster. Moreover, path 3 and 16 resulted to be analogous for compound 1 (Fréchet distance equal to 1.81) and similar considerations holds for paths 14 and 18 (Fréchet distance equal to 1.38). This similarity could be observed in the low-D space (pathways 3 and 16 in orange scale, pathways 4 and 18 in blue scale). Investigating the configurations sampled by the ligand in these unbinding pathways in the high-D space is a further evidence of these similarities and differences. In particular, in the Figure 20 C-D, the coordinates of the COM of the ligand were traced in the high-D space for both couple of unbinding trajectories 14 and 18, 3 and 16. In Figure 20 C is illustrated that two different configurations sampled along the unbinding process by the compound 1 for each of the two pathways (14 and 18) perfectly overlap. A similar pattern could be detected for trajectories 3 and 16 in the same Figure 20 D. Additionally, by comparing Figure 20 C and Figure 20 D, it can be seen that even though the COM trace in the high-D space is comparable for pathways 3, 16 and 14, 18, these two groups of pathways are very different in the low-D space, demonstrating their mechanistic dissimilarity. Indeed, the ligand in these two clusters of trajectories leaves the binding site in completely different orientations: with the quinazoline ring first in pathways 14 and 18, with the 1,3-dihydroisoindol-2-yl ring first in pathways 3 and 16. Concluding, the low Fréchet distance represents a valid measure to compare and identify high similarities within a set of unbinding trajectories.

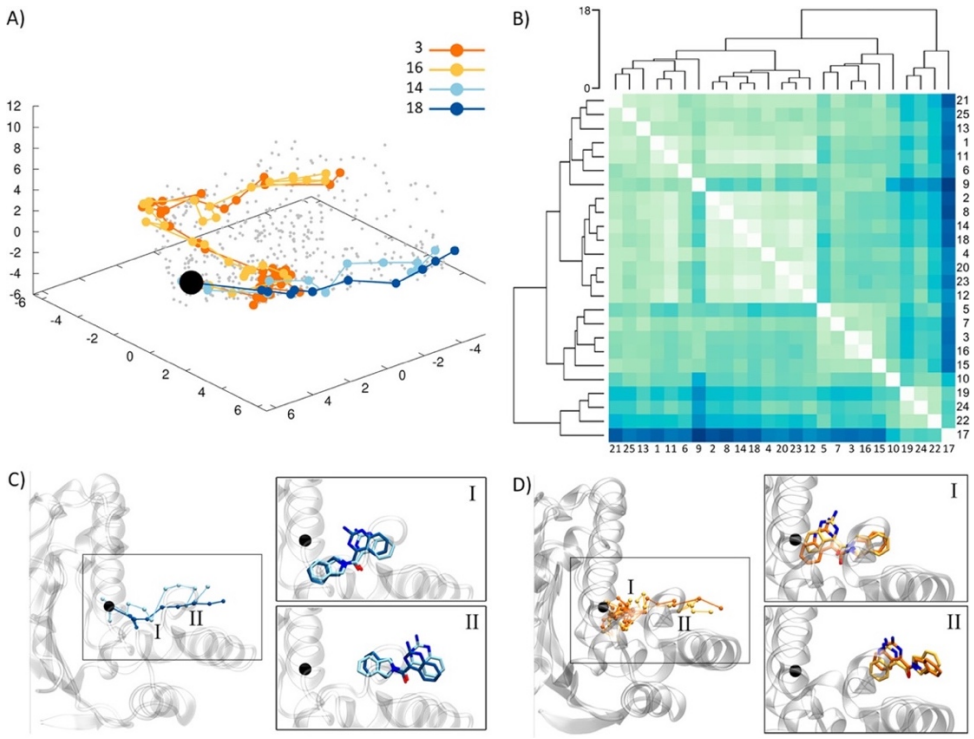


Figure 20. Dimensionality reduction and dendrogram interpretation of the unbinding pathways of compound 1. A) All the ligand configurations (grey points) are represented in the low-D space. Projections of trajectories 3 and 16 and 14 and 18 are identified with different colors: in orange scale and blue scale, respectively. B) Heatmap of the Fréchet distances between all the 25 sMD independent runs colored according the similarity level. In particular, high similarity (low values) is indicated in white, while low similarity (high values) is indicated in blue. The result of the hierarchical cluster analysis is reported on the top and on the left side of the heatmap in form of a dendrogram. C) In left panel are reported the ligand COM coordinates along the unbinding pathway in the high-D space from the initial bound pose (black dot) to the bulk solvent. In right panel are represented two sample configurations visited during the unbinding process in pathways 14 and 18. D) Same description depicted for section C, but relative to trajectories 3 and 16.

3.2.4 Conclusions

Here, we investigated the efficiency of sMD simulations to predict the residence time for a series of structurally different HSP90α inhibitors. Particularly, the experimental residence times of this series of compounds cover a wide-range of 3 order of magnitude (from 2 s up to almost 1 h). First of all, we were able to rank in a reliable way the investigated series of compounds, apart from one inhibitor that emerged to be an outlier of our procedure. Statistical significance and robustness of the observed events (25 sMD simulations for each inhibitors) was evaluated by bootstrapping analysis. Moreover, we demonstrated that even little changes in chemical structure can be correctly ranked according to the experimental evidence. For instance, the two structurally related

compounds 4 and 5, in which to the indazole scaffold is added a piperidine substituent in the former and a morpholine group in the latter, present a different affinity of 1 order of magnitude. Secondly, we developed a fully automated data analysis approach in order to easily identify similar features on the ensemble of unbinding pathways. Despite the quite low scaling factor of 0.45 here used, the procedure seems to be encouraging since we were able to unambiguously identify similar pathways in the high-D space. An interesting aspect that could be investigated in the future, at more conservative scaling factors, is the correlation between unbinding routes and computational residence times. Indeed, it is reasonable to expect that, as long as the kinetic model is capable to properly predict the residence times for the investigated compounds, some features reflecting the mechanistic aspect of the unbinding pathway could be derived. Moreover, this piece of information could be further exploited to guide other sampling strategies in order to identify minimum free energy paths and the related potentials of mean force since the pathways described in the low-D space can be directly used as PCVs on which perform biased-MD simulations. This would lead to a better characterization of the ligand unbinding process in qualitative but also quantitative terms. In particular, it would be possible to identify the relevant metastable states visited by the ligand along the unbinding path and quantify the free energy.

3.3 A Semi-automatic protocol to identify path collective variables to perform well-tempered metadynamics and to compute the protein-ligand binding potential of mean force.

3.3.1 Introduction

Of fundamental importance in the drug discovery field is the identification of small molecule ligands that are able to bind and modulate the activity of biological macromolecules involved in pathological conditions. In the simplest and general description, binding of small molecules to biological target is described, most of times, as a two-state, all-or-none process in which the ligand is either free in the solvent or locked in the binding site. However, an accurate characterization of the thermodynamics and kinetics leading the formation of complexes is required to deeply apprehend the molecular principles driving all biological interactions. In particular, thermodynamics and kinetics the properties of any protein-ligand complex are a function of the potential energy landscape through an averaging process that depends on the statistical mechanics ensemble used. Even if thermodynamics and kinetics are usually described separately, as widely discussed in the Introduction section, they are both linked to free energy. In particular, thermodynamics provides the driving force of protein-ligand binding that is identified by the binding free energy or the equilibrium dissociation constant. The protein-ligand complex is stably formed only when the change in Gibbs free energy (binding free energy), ΔG_b, is negative at equilibrium conditions. Because the protein-ligand association is quantified in terms of the magnitude of the negative ΔG_b, this reflects the stability of the protein-ligand system or of the binding affinity between the interacting partners, typically referred to K_d. Kinetic rates, namely k_{on} and k_{off}, depend respectively on the free energy difference between the more stable bound and unbound states and the transition states, which are complicated to observe and quantify computationally. Nowadays, kinetic and thermodynamic parameters describing the protein-ligand binding process are gaining growing importance as a fundamental information to keep in account in drug discovery, especially in the lead optimization phase, where a rational optimization of these quantities is desired in order to improve the binding proprieties.[5,189] Consequently, an accurate prediction of the free energy profile associated to the protein-ligand binding event could be of a certain importance for modern drug design.

Nowadays, several different biophysical experimental techniques are able to determine thermodynamic and kinetic quantities, but they encounter some difficulties in characterizing in detail complex reaction pathways, especially when several different metastable states are involved in the process. In this context, computer simulations, especially MD, could be extremely helpful since an all-atom description of the protein-ligand binding event is provided. Moreover, in the limit

of an ergodic simulation, information about metastable states and also about transition states can be evaluated and the related affinity and rates would be calculated. However, the main drawback of plain MD simulations is related to the high computational resources needed to simulate such rare event as protein-ligand binding with an adequate statistic. Because of these difficulties is still impossible to routinely use plain MD in the drug discovery pipeline. Among all the enhanced sampling methods, which have been developed in order to provide a valid alternative to simple plain MD,[67] MetaD[70] allows an efficient exploration of the phase space guiding the sampling along reaction coordinates (CVs). However, identifying a proper set of CVs is a far from trivial task and the efforts required become greater as the complexity of the investigated process increases.[76] Moreover, this aspect is further exacerbated in the case of the molecular recognition process, in which many degrees of freedom are involved. In this context, the simplest way to detect a proper set of CVs is to proceed by trial and errors, changing set of CVs in several different MetaD simulations and evaluating a posteriori the suitable combination of variables. This way of proceeding is rather inefficient and prevents any technological diffusion of the approach. The PCVs has been specifically introduced to manage complex reaction pathways such as protein-ligand binding since it is possible to describe the whole process in a configurational space from the initial to the final states introducing only two variables that describe the position of the system along the preassigned path and the distance from it.[79] The basic idea is to guide MetaD sampling along a preassigned pathway, which represents the progression along the ligand binding process, while exploring adjacent regions of the phase space that are orthogonal to the initial guess path.[190–192] Even though multiple pathways can be contemplated, the most optimal path, the one in which the minimum free energy lies, is unique. From this path, it is possible to know the detail information about the binding process. Although being important, finding an optimal path with a high reaction rate or low free energy barriers is far from trivial.

Herein, we present our semi-automated strategy to accelerate the construction and the optimization process of the guess path in order to estimate the free energy along it. The guess path is obtained in two main steps. First, a sequence of waypoints along the path is placed by a machine learning algorithm, introduced by Ferrarotti et al. in 2018,[49] producing a smooth string known as principal path, including relevant intermediate states visited by the ligand along its way to the binding site, from an initial noisy binding trajectory. Second, an optimization protocol requiring a series of subsequent SMD simulations is applied to uniform the spacing between consecutive frames in the path, which is a necessary technical condition to apply the PCVs.[79] Taking advantage of such guess path, the FES explored by the ligand during the dissociation event could be reconstructed through PCVs-based well-tempered MetaD. By setting a wall on the $z(x)$ variable, we let the ligand free to explore several different routes to access to the binding site inside a prescribed phase space.

Because of this freedom, we might be able to identify the path closest to the minimum free energy path that could be characterized in terms of its related intermediate metastable and transition states. Despite the efficiency and its ability of sMD in reproducing the experimental ligand binding mode and accurately predicting relative off-rates of protein-ligand complexes, sMD trajectories, obtained using a scaling factor lower than 0.5, could not be exploited to extract reliable pathways close to the minimum free energy path since the mechanistic and energetic details are heavily approximated. This because, in sMD the ligand follows partially unphysical (un)binding routes as the simulation are performed at high temperature (lower scaling factor). Thus, in order to efficiently reproduce the protein-ligand binding event we exploited the MD-Binding approach, saving computational time while preserving the accuracy.[83] Using electrostatics as the bias to induce protein-ligand recognition and binding via an adaptive behavior, the binding trajectory generated by MD-Binding approach are expected to be potentially comparable to plain MD simulations. In particular, showing mechanistic commonalities with the trajectories obtained via plain MD simulations, the ligand, during a MD-Binding simulation, could follow physical binding routes close to the minimum free energy path allowing our protocol to correctly identify a guess path on which reconstruct a reliable FES.

Here, in the present work we considered a reliable MD-Binding trajectory of an ATP-competitive inhibitor of GSK3β to construct a putative guess path on which we are running a PCVs well-tempered MetaD simulation. In the results and discussion section, we describe the protocol applied to reconstruct the guess path. In addition, some preliminary results of the well-tempered MetaD simulation are presented and the assessment of the convergence is deeply discussed.

3.3.2 Methods

3.3.2.1 Binding trajectory generation

In this work, we considered the ATP-competitive inhibitor of GSK3β 7YG (3-amino-6-(4-{[2-(dimethylamino)ethil]sulfomoyl}phenyl)-n-pyridin-3-ylpyrazine-2-carboxamide), whose experimental binding affinity was available in literature.[156]

The source PDB was 4ACC and only the chain A retained; all the processing was done through BiKi Life Sciences 1.3.5. The ligand was parameterized via RESP charges. The attraction group from the protein side was: "((resname ILE and resid 28) or (resname ASN and resid 30) or (resname VAL and resid 36) or (resname LYS and resid 51) or (resname VAL and resid 76) or (resname LEU and resid 98) or (resname ASP and resid 99) or (resname TYR and resid 100) or (resname VAL and resid 101) or (resname THR and resid 104) or (resname LEU and resid 154) or (resname ASP and resid 166)) and not hydrogen " according to the box renumbering (starting from residue

1). The swich-off selection was "(resname LEU and resid 98) and backbone". The system was simulated for 20 ns and 20 replicas were run.

3.3.2.2 Principal paths in data space by regularized kernel k-means

The principal path, as formulated by Ferrarotti et al.[49], connects two points, defined a priori, in data space and tries to pass through the local mid of the data distribution, capturing the most abstract morphing path between these points. Although this algorithm is formally a regularized version of the k-means clustering algorithm (described in Methods section), its purpose is completely different. This algorithm is based on the idea that, given a dataset and two reference points, w_0 , $w_{N_{c+1}}$, in a vector space $X \in R^d$ (where $d \in N^+$), representing the starting and ending points of a path, a smooth string between them can be found by exploiting the data distribution. The final path identified is different from the well-known definition of shortest path introduced by Dijkstra since the concept of the smoothness is not considered.[193] In particular, the principal path concept is more similar to the definition of the minimum free energy path in the field of statistical mechanics, which is minimal in energetic terms but not automatically in terms of Euclidean distance. Formally, if we consider a set of points $X = \{x_j \in R^d\}$, $j = 1, \dots, N$ and two points w_0, $w_{N_c+1} \in R^d$, the path connecting these two points is defined as an ordered set W of N_c waypoints $w \in R^d$. The principal path is found by minimizing the standard k-means cost function with the addition of a quadratic regularization term that restrains the distance between consecutive waypoints and controls the level of smoothness of the path. The cost function is formalized as:

$$\min_W \frac{\gamma}{2} \sum_{i=1}^{N} \sum_{j=1}^{N_c} \left\| x_i - w_j \right\|^2 \delta(u_i, j) + \frac{\lambda}{2} \sum_{i=0}^{N_c} \| w_{i+1} - w_i \|^2 \quad (40)$$

Where $\delta(u_i, j)$ is the Kronecker delta that gives the membership of the ith sample to jth cluster/waypoint. The first term coincides with the standard k-means cost function and the second term introduces a set of harmonic restrains applied to consecutive waypoints. The two hyperparameters γ, λ regulate the trade-off between data fitting and smoothness of the path as shown in Figure 21.

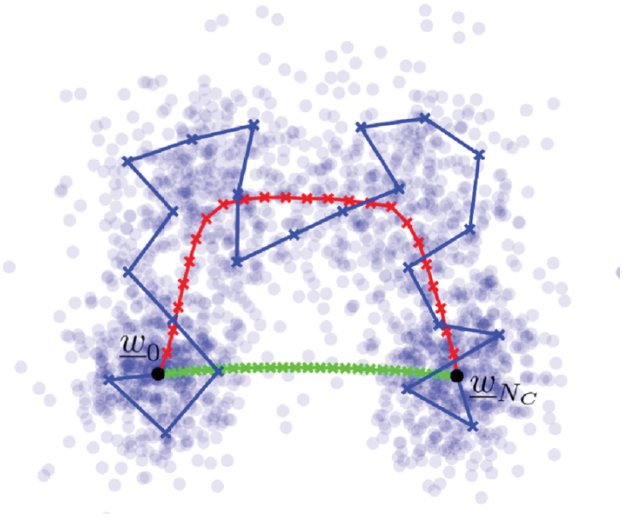


Figure 21. In figure is depicted the hyper-parameter influence in the shape of the path. Data are represented as blue dots, and waypoints as crosses. The irregular blue path represents a configuration where γ is prominent showing an overfitting effect. The green path refers to a configuration where λ is prominent. The red path refers the principal path whit the right trade-off between γ and λ.

3.3.2.3 Equidistant waypoints algorithm

When applied to MD trajectory data in order to identify those configurations that define the PCVs, the path finding algorithm identifies the real frames of the original simulation closest to the calculated principal path, thus making the distance (RMSD) between neighbouring frames not uniform. As reported in the original paper introducing PVCs, consecutive frames have to be as equidistant as possible to ensure a proper mapping between the formal variable s and the underlying metric, especially when a bias simulation is performed.[79] From this standpoint, the states obtained from the path finding algorithm were not well-suited to provide a functional PCVs, since the distribution of the configurations is not equally distributed in space (Figure 21). Consequently, we devised a protocol based on steered MD simulations[128,194] to uniform the spacing in terms of RMSD between pairs of successive windows by placing additional and equidistant configurations according to the need (Algorithm 1). The RMSD was computed using PLUMED 2.5[151] on the ligand coordinates and the heavy atoms of the protein residues located within a distance of 4 Å from the ligand in the binding site in the bound state. The system was aligned on a selection of 25 C_α atoms that resulted to be particularly stable during a plain MD simulation and uniformly distributed on the protein structure. The target RMSD threshold between consecutive frames along the path was set equal to 1 Å.

Algorithm 1 Equidistant waypoints algorithm

Input
θ_0 initial path
t distance threshold between consecutive frames

Output
θ_n final path

Function equidistant_waypoints_algorithm(θ_0,t)
$\alpha = \theta_0[0]$
$\theta_n = []$
$\lambda = \alpha$

for a in range(0,$|\theta_0| - 2$):
$\beta = \theta_0[a+1]$
$\gamma = \theta_0[a+2]$
$\sigma = \text{RMSD}(\lambda, \beta)$

while σ < t:
a = a+1
$\beta = \theta_0[a+1]$
$\gamma = \theta_0[a+2]$
$\sigma = \text{RMSD}(\lambda, \beta)$
$g = \sigma/t$

while $g \geq 1$:
$X = \text{SMD}(\lambda, \beta, \gamma, t)$
$r, \varphi = \text{RMSD\_nearest}(\lambda, X, t)$

if $|r - t| < 0.001$:
$\theta_n.append(\varphi)$
$\lambda = \varphi$
$g = g - 1$

else:
while $|r - t| > 0.001$:
$X_t = \text{continue\_SMD}(\lambda, \beta, \gamma, t)$
$r, \varphi = \text{RMSD\_nearest}(\lambda, X_t, \beta)$

$\theta_n.append(\varphi)$
$\lambda = \varphi$
$g = g - 1$

SMD = SMD simulation to drag β to a RMSD distance of t from λ, while preserving the RMSD distance from γ.

Continue_SMD = extend the previous SMD simulation.

RMSD_nearest = from the SMD trajectory (X), return the RMSD (r) of each frame (φ).

3.3.2.4 Free energy profile and binding free energy

The GSK-3β protein and ions were modelled according to the Amber ff14.[195] The ligand 7YG was parametrized according to the general Amber force field (GAFF).[145] Partial charges were assigned through RESP procedure.[149] In particular, geometrical optimization and further calculation of the electrostatic potential were carried out via quantum mechanics using the 6-31G* basis set at the Hartree-Fock level of theory. The TIP3P model was used for treating water[141] and an appropriate number of ions were added to neutralize the system. The simulation was performed using the 2016.5 version of the GROMACS MD simulation engine[196] patched with the 2.5 version of PLUMED.[151] Production runs were performed in NVT statistical ensemble using a 2 fs time-step. A cut-off of 12 Å was used for non-bonded interactions, while long-range electrostatics was treated with the Particle-Mesh Ewald scheme employing a grid spacing of 1.2 Å.[62] The temperature of 300 K was controlled using the V-rescale thermostat while bond lengths for chemical bonds involving hydrogens were restrained to their equilibrium values with the LINCS algorithm.[144]

Gaussians with a nominal height of 0.2 kcal/mol were used together with a bias factor of 15 units. The width of the Gaussians was set to 0.2 and 0.01 nm^2 along s and z, respectively, and the space along the z dimension was limited to 0.05 nm^2. The Gaussians deposition time was set to 500 molecular dynamics steps. Having optimized the RMSD between pairs of consecutive frames to 0.1 nm (1 Å), the λ parameter determining the smooth description of the path, was set to 230.0 according to equation (34).

As described by Doudou S. et al,[197] according to statistical mechanics the standard free energy change of binding ΔG_b° between the bound and unbound states is given by:

$$\Delta G_b^\circ = -RTln\left(\frac{Q_{site}}{Q_{bulk}}\right) - RTln\left(\frac{V_{bulk}}{V^\circ}\right) \quad (42)$$

Where Q_{site} and Q_{bulk} denote the partition functions for the bound and unbound regions, respectively, V_{bulk} is the sampled unbound volume and V° is the standard-state volume. Their ratio is computed by integrating the potential of mean force in the site and bulk region respectively via:

$$\frac{Q_{site}}{Q_{bulk}} = \frac{\int_{site} exp\left(-\frac{F(s,z)}{RT}\right)dsdz}{\int_{bulk} exp\left(-\frac{F(s,z)}{RT}\right)dsdz} \quad (43)$$

Where $F(s,z)$ is the FES as reconstructed by well-tempered MetaD as a function of *s(x)* and *z(x)* and defined to be zero at its lowest point.

3.3.3 Results and Discussion

3.3.3.1 MD-Binding simulation

As already introduced, in order to identify a putative guess path close to the minimum free energy path along which to reconstruct the binding free energy, the MD-binding approach within the BiKi suite was applied to investigate protein-ligand association event of the ligand 7YG to GSK3β. Over 20 replicas we selected the one which exhibited the lower RMSD with respect to the co-crystal structure.

The final MD-binding trajectory was first post-processed in order to discard all the configurations that are not relevant for the association process under investigation. In particular, we kept into account only the part of the trajectory that ranges from the frame in which the ligand is correctly docked into the binding site to the frame in which the ligand is completely solvated. The solvated state was identified as the first frame in which the ligand lost all the contacts with the protein within a distance of 6 Å.

3.3.3.2 Path generation and optimization

In order to calculate the principal path with the regularized k-means algorithm, an input $N \times d$ matrix need to be calculated. In this work, we defined the $N \times d$ matrix as the coordinates of the heavy atoms of the ligand and the protein residues located within a distance of 4 Å from the ligand in the binding site in the bound state sampled during the post-processed trajectory, X. In particular, in every row of X are reported the coordinates in 3D-dimensional space of all the heavy atoms of the ligand and the residues of the protein (d) in a specific frame of the trajectory (N is the total number of frames).
By applying the path finding algorithm, we detect the milestone frames that define a smooth string of real coordinates from the initial noisy trajectory. The smoothing parameter, calculated as the ratio between γ and λ (s = γ/λ), used to identify the principal path was set to 3500 as shown in Figure 22 and defined via visual inspection of the projected path.

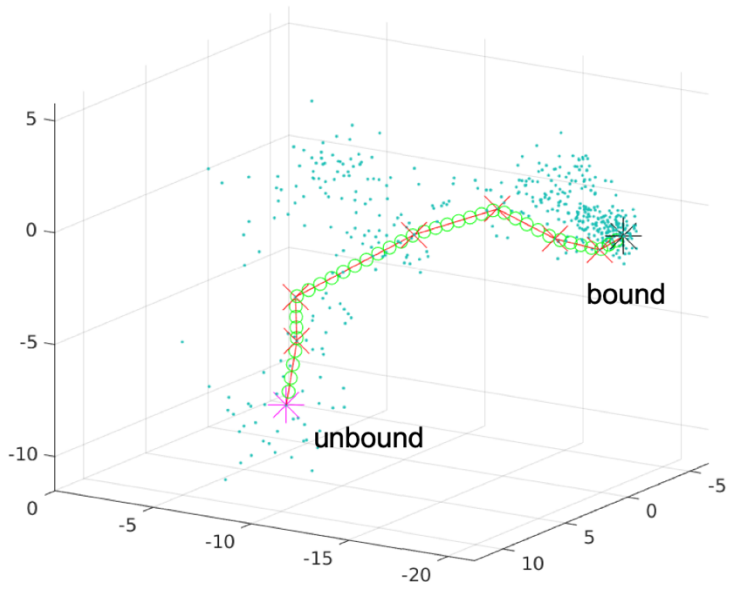


Figure 22. Principal path resulted considering a smoothing factor of 3500. Configurations of the ligand and residues of protein composing the binding site (residues at 4 Å from the ligand in bound state) are represented in the low-dimensional space as blue dots. Configurations that lie on the principal path are represented as green circle. The bound and unbound states are represented as star with different colours.

Once the configurations of the initial trajectory that define the preliminary principal path were detected, it is necessary to perform the optimization step in order to obtain a PCVs in which the distance between consecutive frames is constant. As already described in Methods of this Chapter, this step is achieved by exploiting a series of SMD simulations between each pair of the initial configurations in order to identify those configurations that are at specific distance in terms of RMSD along the original path. In this specific case, the principal path algorithm identified 70 configurations of the original MD-Binding trajectory to define the preliminary path. As shown in Figure 23 left panel, the RMSD between neighbour frames was not constant ranging from about 0.8 to over 4 Å. After the optimization phase, the final PCVs is composed of 102 configurations with a constant RMSD of 1 Å between adjacent frames (Figure 23, right panel).

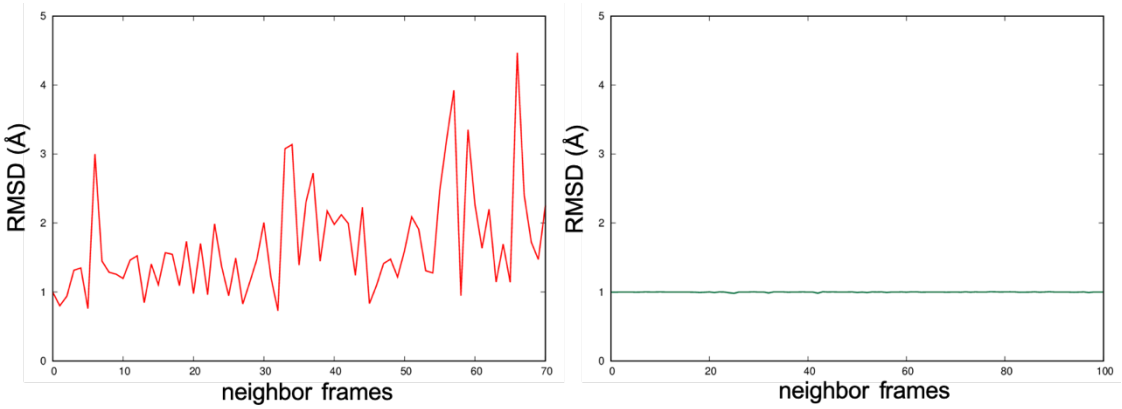


Figure 23. RMSD between adjacent frames before (left) and after (right) the application of the sMD based optimization protocol. In the top panels, the variation of the RMSD along the *s*(x) is reported.

3.3.3.3 Well-tempered MetaD simulations

Well-tempered MetaD was initialized from the bound state of the protein-ligand complex. The following results refer to the statistics collected during the first microsecond of simulation.

The convergence of well-tempered MetaD with PCVs is commonly assessed by checking the achievement of the diffusive behavior of the system along the PCVs in addition to the height of the Gaussian hills deposited. Reaching convergence is necessary to reconstruct a reliable FES by ensuring the complete exploration of the configurational space along the PCVs by the system. However, in real case studies such as protein-ligand binding, reaching convergence is far from a trivial task, because of the high number of slow degrees of freedom involved in the process.

In Figure 24 A, we report the evolution of the system along the *s* variable during the simulation time (i.e. 1 microsecond). As transitory time, we considered the first 350 ns during which the system moved from the bound state (s=102) to the bulk at least twice, thus exploring all the *s* space along the path.

Well-tempered MetaD is supposed to be converged when the Gaussian heights deposited are almost zero.

In our case, after the equilibration time, the system reached an almost diffusive regime along *s* (Figure 24 A) whereas the Gaussians are continuously changing while progressively decreasing in time (Figure 24 B). As a consequence, the convergence of well-tempered MetaD cannot be assess by simply considering the Gaussian heights or the diffusivity along *s* separately. To support this hypothesis, we reported 3 representative one-dimensional profiles of the free energy at different simulation times obtained by integrating over *z(x)* variable and projected along the *s(x)* variable by using the standard plumed tools (Figure 24 C).

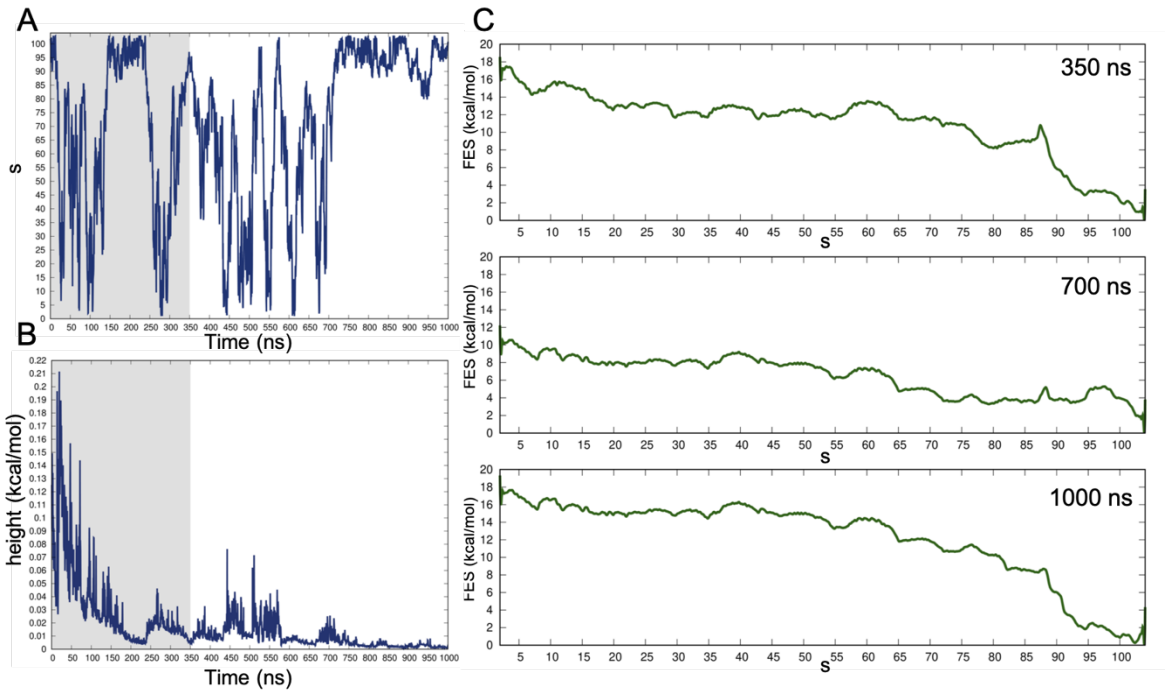


Figure 24. Preliminary results of well-tempered MetaD of the first 1 μs of simulation. A) System progression along the s variable of PCVs in function of time. B) Gaussian heights in function of time. C) One-dimensional profiles of the free energy at different times (350 ns, 700 ns, 1000ns), obtained by integrating over *z* variable and projected along *s*.

By looking at the first profile, we noticed that immediately after the equilibration time (350 ns) the free-energy difference between bound and unbound states results to be between 14-18 kcal/mol.
The second profile was computed when the system reaches the diffusive regime along s fulfilling a necessary precondition for MetaD convergence. However, Gaussian heights are not stable to almost zero value suggesting that the system is still sampling the potential energy surface.
Conversely, by proceeding with the statistics until 1 microsecond, the diffusive regime along s is lost, due to the ligand stucked into the binding site, but the Gaussian heights progressively decrease the zero value, which is another necessary precondition for assessing the convergence of well-tempered MetaD. In this last case, it is interesting to notice how the free energy difference between the bound and unbound states deviates from the experimental value.
Thus, from these considerations, we find that evaluating the Gaussian heights and the diffusivity along the collective variable are both necessary preconditions to assess the convergence of well-tempered MetaD simulation, but they are not sufficient when considered alone as done in other works.[191] However, meeting these two conditions in a real case study requires a huge computational effort and the deep knowledge of the slow degrees of freedom involved in the process under investigation, a not realistic scenario.

From this point, we proceed our study trying to find an alternative way to determine a reliable binding free energy with the available data.

In order to correctly calculate the binding free energy ΔG_b° according to equation (42), it is necessary to identify those configurations that refer to the unbound state and those that could be ascribed to the bound state of the ligand. When the PCVs are considered, this objective results in finding an optimal value of the *s* variable able to discriminate between the docked and solvated states of the ligand.

With the aim of understanding the robustness of free energy estimates we quantitatively evaluated how the ΔG_b° changes as a function of the s value chosen to compute the partition functions in equation (43). For simplicity, we avoided to include the second term of equation (42), relative to the ratio of sampled volume to the standard volume, since it represents a small contribution to the total ΔG_b°. As expected, the profile reaches a plateau region (i.e. almost constant ΔG_b°) in correspondence with the complete solvated state of the ligand due to the fact that these configurations are energetically comparable (s between 10-50). As shown in Figure 25, where profiles at different simulation times after the equilibration phase are reported, the trend of all the curves is consistent, in particular in the s region corresponding to the unbound state. This evidence could stand for the fact that the system has completely explored the path along s at least once.

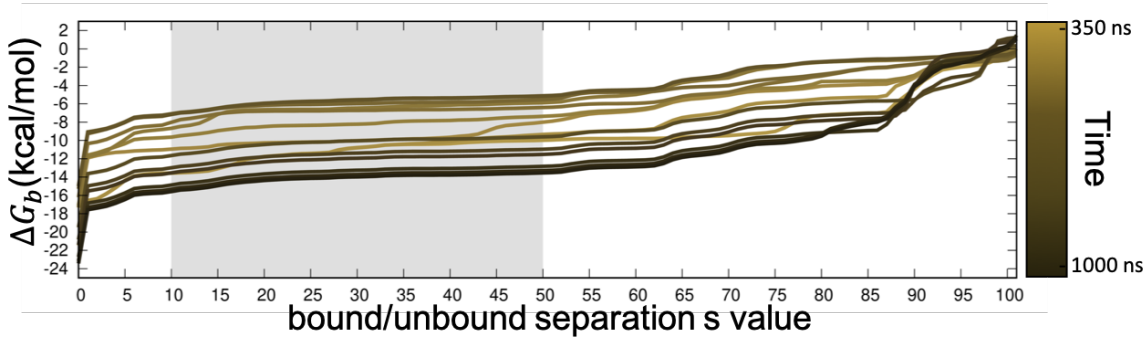


Figure 25. Profiles at different simulation times (between 350 ns to 1 μs) of the ΔG_b calculated as a function of s value chosen to define the unbound and bound region along the path.

To support this analysis and to better characterize the optimal s frame discriminating the bound/unbound states, we also evaluated the solvation shell around the ligand during the binding path (Figure 26, above panel). According to this profile, it is possible to identify three main regions corresponding to the bound state (s frames between 85 and 102), pre-bound state (s frames 50-85) and the unbound state (s frames 1-50). As expected, there is a clear consistency between the profiles reported in Figure 25 and the unbound region identified observing the water shell around the ligand. Thus, we are quite confident to identify in s = 50 to be a reasonable threshold value to discriminate

the bound and unbound states of the ligand, which corresponds to the separation between the completely unbound and the pre-bound state (Figure 26).

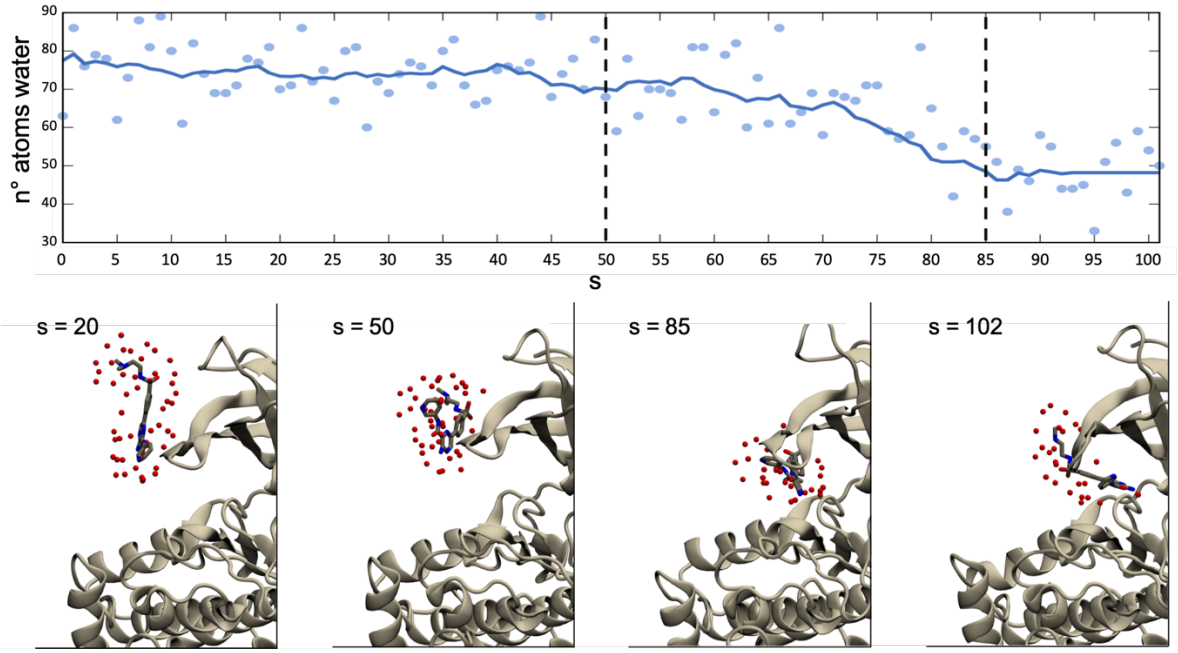


Figure 26. Qualitative evaluation of the water shell surrounding the ligand along s variable of PCVs. In the above panel, are reported the number of atoms water at 3 Å of the ligand for each s frame of the PCVs. In the bottom panel, different configurations of the system are reported to better visualize relevant frames that identify the three main regions: completely bound state (s = 102), frame separating the bound to prebound state (s = 85), frame separating the prebound to completely unbound state (s = 50), and a chosen configuration to represent the completely unbound state (s = 20).

By comparing these profiles at difference times (Figure 25), we noticed that the computed ΔG_b° is sampling a range of values around the experimental value without gradually converging to it. This is another confirmation that the convergence is still not achieved.

However, with the available data, an estimation of the ΔG_b° can be given by considering the average ΔG_b° at s = 50 computed on the free energy profiles at different simulation times which profile is reported in Figure 27. Thus, by applying this procedure, the binding free energy, ΔG_b°, results to be 8.6 kcal/mol, which differs from the experimental value by 2 kcal/mol.

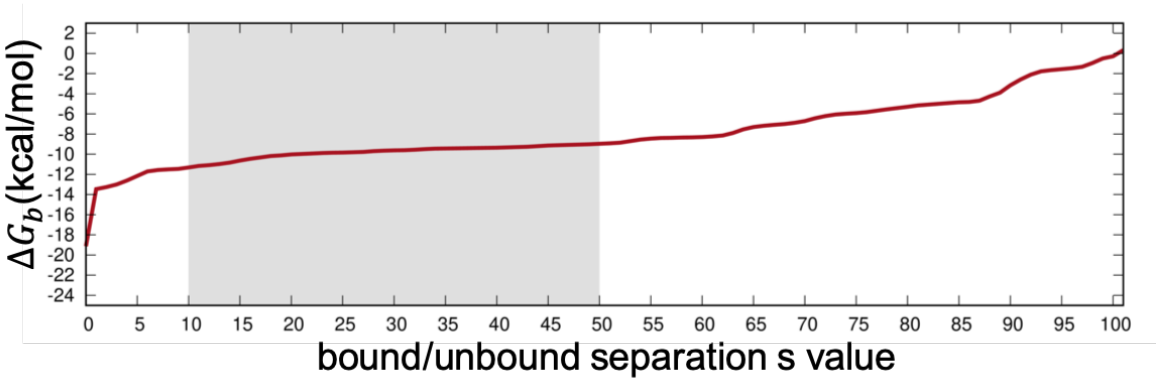


Figure 27. Average of different profiles at simulation time between 350 ns to 1 μs of the ΔG_b calculated as a function of s value chosen to define the unbound and bound region along the path.

In conclusion, we would like to observe some summarizing points:

- Well-tempered metadynamics on the *s* and *z* variables in a protein-ligand binding problem does not exhibit any convergence in 1 us
- This may be due to a sub-optimal definition of the *s* variable, but it represents a realistic scenario in which the a priori knowledge on the system is limited

Despite the lack of convergence, it is interesting to note that the MetaD free energy results oscillates around the true experimental value as time proceed. This means that either the collective variable is not optimal or the well-tempered mechanism, despite the theory, does not converge to a constant FES but instead oscillates around the expected value.

Despite the well-tempered MetaD simulation has not reached the expected convergence and the system is still oversampling the bound state thus leading to a continuously decreasing ΔG_b°, we think that taking into account the average of the ΔG_b° at different simulation times might be an acceptable approximation of the ΔG_b° value at a converged regime in realistic scenario where the domain knowledge and the computing time is limited.

3.3.4 Conclusion

In this study, we introduced a semi-automatic protocol to generate an optimal "guess path" which can be exploited as PCVs by enhanced sampling techniques to guide the sampling along and around it. This "guess path" is essentially a set of consecutive frames of the system captured at different time steps along the process under investigation.

Herein, we took advantage of this semi-automatic protocol to construct a guess path for the inhibitor 7YG binding to GSK3β, in order to exploit it for subsequent PCVs well-tempered MetaD

simulation. The associated FES reconstructed as a negative image of the Gaussians deposited during time, would allow calculating the ΔG_b°.

To generate the path, we perform a preliminary MD-Binding simulation in order to observe the binding event at least once within a reasonable time. Basing on the resulting trajectory, the principal path algorithm introduced by Ferrarotti et al.[49] was applied to identify a prior guess path which pass through the local distribution of whole set of configurations. The resulting frameset is then post-processed through a series of short consecutive SMD simulations to ensure a uniform distance, in terms of RMSD, between adjacent frames.

Although the results of the PCVs well-tempered MetaD simulation here reported are only preliminary, we tried to interpret them in a rigorous way in order to identify some general key aspects that could be used to predict a reliable ΔG_b° even if the convergence is not reached.

4 CONCLUSIONS

Characterizing the molecular recognition process is of utmost importance for drug discovery since it is involved in many biological processes with consequent pharmaceutical implications. In particular, a deep knowledge of protein-ligand interactions is essential for a successful drug design. Capturing the whole binding process on a mechanistic level could be much interesting since, ideally, every contribution to the establishment of the molecular complex could be exploited in the development of new drugs.

Thus, significant efforts have been made over the years into the investigation of the protein-ligand binding process leading to the development of both experimental and computational techniques able to quantify kinetics and thermodynamics of bio-molecular complexes.

Several experimental techniques such as ITC, NMR and SPR demonstrated to be efficient in studying and quantifying both thermodynamic and kinetic observables. However, these are sometimes expensive and time-consuming. Moreover, a detailed atomic-level description of the complex reaction pathways, with the characterization of all the metastable states is not possible since only a quantitative estimation is allowed.

From the computational point of view, despite multiple successes of molecular docking in predicting the structures of many protein-ligand complexes, the dynamics of binding have been ignored by computational investigations. However, during last decade, there has been growing interest in developing models describing the entire dynamic process upon complex formation, elucidating how ligand binding can modify the structure and function of the biological target which is of fundamental importance to correctly interpret the binding event.

In this perspective, MD has been increasingly considered as the computational method of choice to investigate the entire protein-ligand binding process and, in theory, estimate the related thermodynamic and kinetic observables. However, due to the significant computational effort involved in a plain MD protocol, achieving an accurate and comprehensive description of the process and the related estimation of thermodynamic and kinetic parameters for larger sets of compounds is not feasible yet.

Therefore, different enhanced sampling strategies have been developed in order to overcome the limitations of plain MD, mostly due to the timescales problem, and to allow a complete characterization of the investigated process.

Here, we devised three different protocols according to the specific aspect of protein-ligand binding process that we were interested in: dynamical docking, estimation of kinetic observables and reconstruction of the free energy surface associated to the binding process.

First of all, we developed a dynamic docking protocol based on sMD simulations, in which the protein is let completely flexible in order to predict the protein ligand binding pose with a reasonable computational effort. Briefly, in this semi-automatic protocol, the ligand is initially placed in the bulk in a random configuration, at the basis of a cylindrical volumetric confinement which extends from the binding pocket to the solvent. A series of ten independent sMD was performed and, through a cluster analysis of all the metastable states visited along the trajectory, we were able to successfully identify the experimental binding mode among the top ranked clusters. In particular, we tested our approach on the prediction of a series of five different compounds binding two relevant pharmacological protein; the GSK3β and the N-terminal domain of HSP90α, for which the experimental crystal structures were available.

Provided the importance of leveraging kinetic observables in the drug discovery process and the necessity of developing methods to routinely predict drug residence time, in the second application we further investigated the applicability of sMD in describing the kinetic behavior of a series of drug-like molecules. Specifically, we considered a series of congeneric ligand of the N-terminal domain of HSP90α presenting a wide range (from less than 2s up to almost 1h) of experimental residence times. Here, without any a priori information about the unbinding mechanism, we investigated the capability of sMD simulations to estimate the residence times and the unbinding pathways.

We demonstrated that, even if it is difficult to differentiate in single runs structurally similar ligands, we were able to properly rank them, in agreement with experimental results, reaching a statistical significance and robustness by applying bootstrapping analysis.

Moreover, we devised a fully automated method to analyze the unbinding trajectories, detecting similar features. The resulting pathway describing the unbinding process, in the future could be further exploited as a collective variable to guide more accurate biased-MD methods, with the purpose of identifying the minimum free energy path and the related potential of mean force. In this way a more reliable characterization of the mechanism underlying protein-ligand unbinding process, with the quantification of the involved barriers and the identification of the metastable states encountered would be possible.

However, the binding and unbinding routes resulting from sMD simulations could not be considered as reliable pathways close to the minimum free energy path since it is equivalent to perform the simulation at high temperature in which the ligand could explore all the phase space.

In order to extract from a MD simulation a reliable (un)binding path on which to reconstruct the free energy surface and gain into the thermodynamics and kinetics related with the binding event, we took advantage of the MD-Binding method which approaches the accuracy of plain MD.

This approach was devised in the third application, where we applied a machine learning algorithm to extract a principal binding path from a MD-Binding simulation of an ATP-competitive inhibitor of GSK3β. The underlying idea is to guide well-tempered MetaD sampling, a well-established enhanced sampling method, exploring all the metastable states along the principal path identified, and recover the free energy surface related to completely characterize the binding process with the estimation of the affinity in terms of the difference of the standard binding free energy ΔG_b° . Despite only preliminary results of the first microseconds of PCVs well-tempered MetaD simulation are reported, we propose an analysis to anyway try to give an estimate of ΔG_b° even if the convergence is not completely reached. This is particularly important practically as in order to effectively use such tools in real-world drug discovery projects it is necessary to work with a limited amount of computing time. Thus, it is interesting to try to get "numbers" even from not completely rigorous simulations and compare them to experimental values. While this strategy is not scientifically completely rigorous from an engineering and applicative viewpoint is rather useful.

5 REFERENCES

1. Du, X. *et al.* Insights into protein–ligand interactions: Mechanisms, models, and methods. *Int. J. Mol. Sci.* **17**, 1–34 (2016).
2. Copeland, R. A., Pompliano, D. L. & Meek, T. D. Drug–target residence time and its implications for lead optimization. *Nat. Rev. Drug Discov.* **5**, 730–739 (2006).
3. Perozzo, R., Folkers, G. & Scapozza, L. Thermodynamics of protein-ligand interactions: History, presence, and future aspects. *J. Recept. Signal Transduct.* **24**, 1–52 (2004).
4. Bernetti, M., Cavalli, A. & Mollica, L. Protein-ligand (un)binding kinetics as a new paradigm for drug discovery at the crossroad between experiments and modelling. *Medchemcomm* **8**, 534–550 (2017).
5. Deganutti, G. & Moro, S. Estimation of kinetic and thermodynamic ligand-binding parameters using computational strategies. *Futur. Med. Chem.* **9**, 507–523 (2017).
6. Li, H. M., Xie, Y. H., Liu, C. Q. & Liu, S. Q. Physicochemical bases for protein folding, dynamics, and protein-ligand binding. *Sci. China Life Sci.* **57**, 287–302 (2014).
7. Caro, J. A. *et al.* Entropy in molecular recognition by proteins. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 6563–6568 (2017).
8. Chodera, J. D. & Mobley, D. L. Entropy-enthalpy compensation: Role and ramifications in biomolecular ligand recognition and design. *Annu. Rev. Biophys.* **42**, 121–142 (2013).
9. Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der Dtsch. Chem. Gesellschaft* **27**, (1894).
10. Vogt, A. D., Pozzi, N., Chen, Z. & Di Cera, E. Essential role of conformational selection in ligand binding. *Biophys. Chem* 13–21 (2014). doi:10.1016/j.cortex.2009.08.003.Predictive
11. Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci.* **44**, 98–104 (1958).
12. Bosshard, H. R. Molecular recognition by induced fit: How fit is the concept? *News Physiol. Sci.* **16**, 171–173 (2001).
13. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem Sci* **35**, 539–546 (2010).
14. Ma, B., Kumar, S., Tsai, C. J. & Nussinov, R. Folding funnels and binding mechanisms. *Protein Eng.* **12**, 713–720 (1999).
15. Tsai, C.-J., Kumar, S., Ma, B. & Nussinov, R. Folding funnels, binding funnels, and protein function. *Protein Sci.* **8**, 1181–1190 (1999).
16. Tobi, D. & Bahar, I. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 18908–

18913 (2005).

17. Kastritis, P. L. & Bonvin, A. M. J. J. On the binding affinity of macromolecular interactions: Daring to ask why proteins interact. *J. R. Soc. Interface* **10**, (2013).
18. Freire, E., Mayorga, O. L. & Straume, M. Isothermal titration calorimetry. *Anal. Chem.* **62**, 950A-959A (1990).
19. Becker, W., Bhattiprolu, K. C., Gubensäk, N. & Zangger, K. Investigating Protein–Ligand Interactions by Solution Nuclear Magnetic Resonance Spectroscopy. *ChemPhysChem* **19**, 895–906 (2018).
20. de Mol, N. J. & Fischer, M. J. E. Chapter 5. Kinetic and Thermodynamic Analysis of Ligand–Receptor Interactions: SPR Applications in Drug Development. *Handb. Surf. Plasmon Reson.* 123–172 (2008). doi:10.1039/9781847558220-00123
21. Atkins, P., De Paula, J. & Keeler, J. *Atkin's physical chemistry*. (Oxford Univeristy Press, 2018).
22. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. (1982).
23. Talele, T., Khedkar, S. & Rigby, A. Successful Applications of Computer Aided Drug Discovery: Moving Drugs from Concept to the Clinic. *Curr. Top. Med. Chem.* **10**, 127–141 (2010).
24. Carlson, H. A. Protein flexibility and drug design: How to hit a moving target. *Curr. Opin. Chem. Biol.* **6**, 447–452 (2002).
25. Gioia, D., Bertazzo, M., Recanatini, M. & Cavalli, A. Dynamic Docking : A Paradigm Shift in Computational Drug Discovery. *Molecules* **22**, 1–21 (2017).
26. Hu, X., Maffucci, I. & Contini, A. Advances in the treatment of explicit water molecules in docking and binding free energy calculations. *Curr. Med. Chem.* **25**, (2018).
27. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **28**, 849–857 (1985).
28. Amadasi, A. *et al.* Mapping the Energetics of Water–Protein and Water–Ligand Interactions with the "Natural" HINT Forcefield: Predictive Tools for Characterizing the Roles of Water in Biomolecules. *J. Mol. Biol.* **358**, 289–309 (2006).
29. Verdonk, M. L., Cole, J. C. & Taylor, R. SuperStar: A Knowledge-based Approach for Identifying Interaction Sites in Proteins. *J. Mol. Biol.* **289**, 1093–1108 (1999).
30. Michel, J., Tirado-Rives, J. & Jorgensen, W. L. Prediction of the Water Content in Protein Binding Sites. *J. Phys. Chem. B* **113**, 13337–13346 (2009).
31. Young, T., Abel, R., Kim, B., Berne, B. J. & Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 808–813 (2007).

32. Zheng, M., Li, Y., Xiong, B., Jiang, H. & Shen, J. Water PMF for predicting the properties of water molecules in protein binding site. *J. Comput. Chem.* **34**, 583–592 (2013).
33. Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F. & Mason, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model.* **47**, 279–294 (2007).
34. Feixas, F., Lindert, S., Sinko, W. & McCammon, J. A. Exploring the Role of Receptor Flexibility in Structure-Based Drug Discovery. *Biophys. Chemi* **186**, 31–45 (2014).
35. Buonfiglio, R., Recanatini, M. & Masetti, M. Protein Flexibility in Drug Discovery: From Theory to Computation. *ChemMedChem* **10**, 1141–1148 (2015).
36. Craig, I. R., Essex, J. W. & Spiegel, K. Ensemble docking into multiple crystallographically derived protein structures: An evaluation based on the statistical analysis of enrichments. *J. Chem. Inf. Model.* **50**, 511–524 (2010).
37. Bottegoni, G., Rocchia, W., Rueda, M., Abagyan, R. & Cavalli, A. Systematic exploitation of multiple receptor conformations for virtual ligand screening. *PLoS One* **6**, (2011).
38. Sousa, S. F. *et al.* Protein-Ligand Docking in the New Millennium – A Retrospective of 10 Years in the Field. *Curr. Med. Chem.* **20**, 2296–2314 (2013).
39. Rocchia, W. *et al.* Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. *J. Comput. Chem.* **23**, 128–137 (2002).
40. Wang, J., Morin, P., Wang, W. & Kollman, P. A. Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *J. Am. Chem. Soc.* **123**, 5221–5230 (2001).
41. Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. Semianalytical Treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127–6129 (1990).
42. Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **246**, 122–129 (1995).
43. Grinter, S. Z. & Zou, X. Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design. *Molecules* **19**, 10150–10176 (2014).
44. Ain, Q. U., Aleksandrova, A., Roessler, F. D. & Ballester, P. J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **5**, 405–424 (2015).
45. Shen, C. *et al.* From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 1–23 (2019). doi:10.1002/wcms.1429

46. De Vivo, M., Masetti, M., Bottegoni, G. & Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **59**, 4035–4061 (2016).
47. Salmaso, V. & Moro, S. Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. *Front. Pharmacol.* **9**, 1–16 (2018).
48. De Vivo, M. & Cavalli, A. Recent advances in dynamic docking for drug discovery. *WIREs Comput Mol Sci* 1–10 (2017). doi:10.1002/wcms.1320
49. Ferrarotti, M. J., Rocchia, W. & Decherchi, S. Finding Principal Paths in Data Space. *IEEE Trans. Neural Networks Learn. Syst.* 1–14 (2018). doi:10.1109/TNNLS.2018.2884792
50. Manuscript, A. & Nanostructures, S. P. C. NIH Public Access. *Nano* **6**, 2166–2171 (2008).
51. Buch, I., Giorgino, T. & De Fabritiis, G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci.* **108**, 10184–10189 (2011).
52. Shan, Y. *et al.* How Does a Drug Molecule Find Its Target Binding Site ? *J. Am. Chem. Soc.* **133**, 9181 (2011).
53. Dror, R. O. *et al.* Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc. Natl. Acad. Sci.* **108**, 13118–13123 (2011).
54. Rocchia, W., Masetti, M. & Cavalli, A. Chapter 11 Enhanced Sampling Methods in Drug Design. in *Physico-Chemical and Computational Approaches to Drug Discovery* 273–301 (The Royal Society of Chemistry, 2012). doi:10.1039/9781849735377-00273
55. Born, M. ; Oppenheimer, R. Zur quantentheorie der molekeln. *Ann. Phys.* **389**, 457–484 (1927).
56. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
57. Harvey, M. J., Giupponi, G. & De Fabritiis, G. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.* **5**, 1632–1639 (2009).
58. Feenstra, K. A., Hess, B. & Berendsen, H. J. C. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.* **20**, 786–798 (1999).
59. Hunenberger, P. H. Thermostat algorithms for molecular dynamics simulations. in *Advanced in Polymer Science* (ed. Springer, B.) **173**, 105–149 (2005).
60. Frenkel, D.; SMit, B. *Understanding Molecular Simulation*. (Academic Press Inc, 2001).
61. Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*. (1989).
62. Essmann, U. *et al.* A smooth particle mesh Ewald method A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593 (1995).

63. Simonson, T. & Archontis, G. ree E nergy S im ulations C om e of A ge : P rotein - L igand ecognition. **35**, 430–437 (2002).
64. Gilson, M. K., Given, J. A., Bush, B. L. & McCammon, J. A. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J.* **72**, 1047–1069 (1997).
65. Hothersall, J. D., Brown, A. J., Dale, I. & Rawlins, P. Can residence time offer a useful strategy to target agonist drugs for sustained GPCR responses? *Drug Discov. Today* **21**, 90–96 (2016).
66. Copeland, R. A. The drug–target residence time model: a 10-year retrospective. *Nat. Rev. Drug Discov.* **15**, 87 (2015).
67. Abrams, C. & Bussi, G. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy* **16**, 163–199 (2014).
68. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23**, 187–199 (1977).
69. Grubm ller, H., Heymann, B. & Tavan, P. Ligand Binding: Molecular Mechanics Calculation of the Streptavidin-Biotin Rupture Force. *Science (80-.).* **271**, 997–999 (1996).
70. Laio, A. & Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 12562 (2002).
71. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).
72. Hamelberg, D., Mongan, J. & McCammon, J. A. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120**, 11919–11929 (2004).
73. Nakajima, N., Nakamura, H. & Kidera, A. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *J. Phys. Chem. B* **101**, 817–824 (1997).
74. Sinko, W., Miao, Y., de Oliveira, C. A. F. & McCammon, J. A. Population Based Reweighting of Scaled Molecular Dynamics. *J. Phys. Chem. B* **117**, 12759−12768 (2013).
75. Tsujishitaj, H., Moriguchi, I. & Hirono, S. Potential-Scaled Molecular Dynamics and Potential Annealing: Effective Conformational Search Techniques for Biomolecules. *J. Phys. Chem.* **97**, 4416–4420 (1993).
76. Barducci, A., Bonomi, M. & Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 826–843 (2011).
77. Barducci, A., Bussi, G. & Parrinello, M. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **100**, (2008).
78. Piana, S. & Laio, A. A bias-exchange approach to protein folding. *J. Phys. Chem. B* **111**,

4553–4559 (2007).

79. Branduardi, D., Gervasio, F. L. & Parrinello, M. From A to B in free energy space. *J. Chem. Phys.* **126**, (2007).
80. Park, S., Khalili-Araghi, F., Tajkhorshid, E. & Schulten, K. Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *J. Chem. Phys.* **119**, 3559–3566 (2003).
81. Jarzynski, C. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.* **56**, 5018–5035 (1997).
82. Jarzynski, C. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **78**, 2690–2693 (1997).
83. Spitaleri, A., Decherchi, S., Cavalli, A. & Rocchia, W. Fast Dynamic Docking Guided by Adaptive Electrostatic Bias: The MD-Binding Approach. *J. Chem. Theory Comput.* **14**, 1727–1736 (2018).
84. BiKi Technologies s.r.l. BiKi Technologies. Available at: http://www.bikitech.com/.
85. Sheinerman, F. B., Norel, R. & Honig, B. Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153–159 (2000).
86. Laio, A. & Gervasio, F. L. Metadynamics: A method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports Prog. Phys.* **71**, (2008).
87. Shao, J., Tanner, S. W., Thompson, N. & Cheatham, T. E. Clustering Molecular Dynamics Trajectories : 1 . Characterizing the Performance of Different Clustering Algorithms. 2312–2334 (2007).
88. Abramyan, T. M., Snyder, James, A., Thyparambil, A. A., Stuart, S. J. & Latour, R. A. Cluster Analysis of Molecular Simulation Trajectories for Systems where Both Conformation and Orientation of the Sampled States are Important Tigran. *J. Comput. Chem.* **37**, 1973–1982 (2016).
89. Velmurugan, T. & Santhanam, T. Computational complexity between K-means and K-medoids clustering algorithms for normal and uniform distributions of data points. *J. Comput. Sci.* **6**, 363–368 (2010).
90. . G. G. Analysis and Implementation of Modified K-Medoids Algorithm To Increase Scalability and Efficiency for Large Dataset. *Int. J. Res. Eng. Technol.* **03**, 150–153 (2014).
91. Kruse, A. C. *et al.* Structure and Dynamics of the M3 Muscarinic Acetylcholine Receptor. *Nature* **482**, 552–556 (2012).
92. Decherchi, S., Berteotti, A., Bottegoni, G., Rocchia, W. & Cavalli, A. The ligand binding mechanism to purine nucleoside phosphorylase elucidated via molecular dynamics and

machine learning. *Nat. Commun.* **6**, (2015).

93. Ferruz, N., Harvey, M. J., Mestres, J. & De Fabritiis, G. Insights from Fragment Hit Binding Assays by Molecular Simulations. *J. Chem. Inf. Model.* **55**, 2200–2205 (2015).
94. Bisignano, P. *et al.* Kinetic characterization of fragment binding in AmpC β-lactamase by high-throughput molecular simulations. *J. Chem. Inf. Model.* **54**, 362–366 (2014).
95. Dror, R. O. *et al.* Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs. *Nature* **503**, 295 (2013).
96. Pande, V. S. Everything you wanted to ask about Markov Models. **52**, 99–105 (2011).
97. Chodera, J. D. & Noè, F. Markav state models of biomolecular conformational dynamics. *Curr Opin Stuct Biol* 135–144 (2014). doi:10.1038/jid.2014.371
98. Doerr, S. & De Fabritiis, G. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *J. Chem. Theory Comput.* **10**, 2064–2069 (2014).
99. Doerr, S., Harvey, M. J., Noé, F. & De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **12**, 1845–1852 (2016).
100. Ferruz, N., Tresadern, G., Pineda-Lucena, A. & De Fabritiis, G. Multibody cofactor and substrate molecular recognition in the myo-inositol monophosphatase enzyme. *Sci. Rep.* **6**, 1–10 (2016).
101. Stanley, N., Pardo, L. & Fabritiis, G. De. The pathway of ligand entry from the membrane bilayer to a lipid G protein-coupled receptor. *Sci. Rep.* **6**, 1–9 (2016).
102. Sabbadin, D. & Moro, S. Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR-ligand recognition pathway in a nanosecond time scale. *J. Chem. Inf. Model.* **54**, 372–376 (2014).
103. Sabbadin, D., Ciancetta, A., Deganutti, G., Cuzzolin, A. & Moro, S. Exploring the recognition pathway at the human A2A adenosine receptor of the endogenous agonist adenosine using supervised molecular dynamics simulations. *Medchemcomm* **6**, 1081–1085 (2015).
104. Cuzzolin, A. *et al.* Deciphering the Complexity of Ligand-Protein Recognition Pathways Using Supervised Molecular Dynamics (SuMD) Simulations. *J. Chem. Inf. Model.* **56**, 687–705 (2016).
105. Zeller, F., Luitz, M. P., Bomblies, R. & Zacharias, M. Multiscale Simulation of Receptor-Drug Association Kinetics: Application to Neuraminidase Inhibitors. *J. Chem. Theory Comput.* **13**, 5097–5105 (2017).
106. Gervasio, F. L., Laio, A. & Parrinello, M. Flexible docking in solution using metadynamics. *J. Am. Chem. Soc.* **127**, 2600–2607 (2005).
107. Provasi, D., Bortolato, A. & FIlizola, M. Exploring Molecular Mechanism of Ligand Recognition by Opioid Receptors with Metadynamics. *Biochemistry* **48**, 10020–10029 (2009).

108. Limongelli, V., Bonomi, M. & Parrinello, M. Funnel metadynamics as accurate binding free-energy method. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 6358–6363 (2013).
109. Pietrucci, F., Marinelli, F., Carloni, P. & Laio, A. Substrate binding mechanism of HIV-1 protease from explicit-solvent atomistic simulations. *J. Am. Chem. Soc.* **131**, 11811–11818 (2009).
110. Söderhjelm, P., Tribello, G. A. & Parrinello, M. Locating binding poses in protein-ligand systems using reconnaissance metadynamics. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 5170–5175 (2012).
111. Yoshida, K., Yamaguchi, T. & Okamoto, Y. Replica-exchange molecular dynamics simulation of small peptide in water and in ethanol. *Chem. Phys. Lett.* **412**, 280–284 (2005).
112. Ostermeir, K. & Zacharias, M. Advanced replica-exchange sampling to study the flexibility and plasticity of peptides and proteins. *Biochim. Biophys. Acta - Proteins Proteomics* **1834**, 847–853 (2013).
113. Luitz, M. P. & Zacharias, M. Protein-ligand docking using Hamiltonian replica exchange simulations with soft core potentials. *J. Chem. Inf. Model.* **54**, 1669–1675 (2014).
114. Kappel, K., Miao, Y. & McCammon, J. A. Accelerated Molecular Dynamics Simulations of Ligand Binding to a Muscarinic G-protein Coupled Receptor. *Q Rev Biophys* **48**, 479–487 (2015).
115. Kamiya, N., Yonezawa, Y., Nakamura, H. & Higo, H. Protein-inhibitor flexible docking by a multicanonical sampling: Native complex structure with the lowest free energy and a free-energy barrier distinguishing the native complex from the others. *Proteins Struture Funct. Bioinformatic* **70**, 41–53 (2007).
116. Bekker, G. J. *et al.* Accurate Prediction of Complex Structure and Affinity for a Flexible Protein Receptor and Its Inhibitor. *J. Chem. Theory Comput.* **13**, 2389–2399 (2017).
117. Shao, Q. & Zhu, W. Exploring the Ligand Binding/Unbinding Pathway by Selectively Enhanced Sampling of Ligand in a Protein–Ligand Complex. *J. Phys. Chem. B* (2019). doi:10.1021/acs.jpcb.9b05226
118. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe Jr., E. W. Computational methods in drug discovery. *Pharmacol. Rev.* **66**, 334–395 (2014).
119. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004).
120. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins Struct. Funct. Genet.* **52**, 609–623 (2003).
121. Teague, S. J. Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discov.* **2**, 527–541 (2003).
122. Davis, A. M. & Teague, S. J. Hydrogen bonding, hydrophobic interactions, and failure of

the rigid receptor hypothesis. *Angew. Chemie - Int. Ed.* **38**, 736–749 (1999).

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

124. Huang, S.-Y. & Zou, X. inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. *J. Chem. Inf. Model.* **50**, 262–273 (2010).

125. Di Martino, G. P., Masetti, M., Ceccarini, L., Cavalli, A. & Recanatini, M. An automated docking protocol for hERG channel blockers. *J. Chem. Inf. Model.* **53**, 159–175 (2013).

126. Li, D., Liu, M. S., Ji, B., Hwang, K. & Huang, Y. Coarse-grained molecular dynamics of ligands binding into protein: The case of HIV-1 protease inhibitors. *J. Chem. Phys.* **130**, (2009).

127. Durrant, J. D. & McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol.* **9**, 71 (2011).

128. Grubmuller, H., Heymann, B. & Tavan, P. Ligand Binding: Molecular Mechanics Calculation of the Streptavidin-Biotin Rupture Force. *Science (80-.).* **271**, 997–999 (1996).

129. Mark, A. E., van Gunsteren, W. F. & Berendsen, H. J. C. Calculation of relative free energy via indirect pathways. *J. Chem. Phys.* **94**, 3808–3816 (1991).

130. Mazanetz, M. P., Withers, I. M., Laughton, C. A. & Fischer, P. M. Exploiting glycogen synthase kinase 3β flexibility in molecular recognition. *Biochem. Soc. Trans.* **36**, 55–58 (2008).

131. Amaral, M. *et al.* Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. *Nat. Commun.* **8**, (2017).

132. Doble, B. & Woodgett, J. R. Author Manuscript / Manuscrit d ' auteur GSK-3 : tricks of the trade for a multi-tasking kinase. *J. Cell Sci.* **116**, 1175–1186 (2003).

133. Pearl, L. H. & Prodromou, C. Structure and Mechanism of the Hsp90 Molecular Chaperone Machinery. *Annu. Rev. Biochem.* **75**, 271–294 (2006).

134. Decherchi, S. & Rocchia, W. A general and Robust Ray-Casting-Based Algorithm for Triangulating Surfaces at the Nanoscale. *PLoS One* **8**, (2013).

135. Decherchi, S., Bottegoni, G., Spitaleri, A., Rocchia, W. & Cavalli, A. BiKi Life Sciences: A New Suite for Molecular Dynamics and Related Methods in Drug Discovery. *J. Chem. Inf. Model.* **58**, 219–224 (2018).

136. Stebbins, C. E. *et al.* Crystal Structure of an Hsp90–Geldanamycin Complex: Targeting of a Protein Chaperone by an Antitumor Agent. *Cell* **89**, 239–250 (1997).

137. Dajani, R. *et al.* Crystal Structure of Glycogen Synthase Kinase 3: Structural Basis for Phosphate-Primed Substrate Specificity and Autoinhibition. *Cell* **105**, 721–732 (2001).

138. Schrödinger Release 2012. No Title.

139. Van der Spoel, D., Lindhal, E., Hess, B. & Team, and the G. development. No Title. *GROMACS User Man. version 4.6.5* (2013).
140. Lindorff-larsen, K. *et al.* Improved side-chain torsion potentials for the. *Proteins Struct. Funct. Genet.* 1950–1958 (2010). doi:10.1002/prot.22711
141. Jorgensen, W. L., Chandrasekhar, J. & Madura, J. D. Comparison of simple potential functions for simulating liquid water Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
142. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals : A new molecular dynamics method Polymorphic transitions in single crystals : A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182 (1981).
143. Bussi, G. *et al.* Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
144. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS : A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).
145. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
146. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **25**, 247–260 (2006).
147. Case, D. A. *et al.* The amber biomolecular simulations programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).
148. Frisch, M. J. *et al.* No Title. *Gaussian 03, Revis. A.1, Gaussian, Inc., Pittsburgh PA* (2003).
149. Cornell, W. D., Cieplak, P., Bayly, C. I. & Kollman, P. A. RESP Charges. 9620–9631 (1993).
150. Humphrey, W., Dalke, A. & Schulten, K. No Title. *J. Mol. Graph. Model.* **14**, 33–38 (1996).
151. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2 : New feathers for an old bird. *Comput. Phys. Commun.* **185**, 604–613 (2014).
152. Kumar, R. g_distMat. Available at: https://github.com/rjdkmr/g_distMat.
153. R Core Team. No Title. (2013).
154. Hooper, C., Killick, R. & Lovestone, S. The GSK3 hypothesis of Alzheimer's disease. *J. Neurochem.* **104**, 1433–1439 (2008).
155. Mazanetz, M. P., Laughton, C. A. & Fischer, P. M. Investigation of the flexibility of protein kinases implicated in the pathology of Alzheimer's disease. *Molecules* **19**, 9134–9159 (2014).
156. Berg, S. *et al.* Discovery of novel potent and highly selective glycogen synthase kinase-

3beta (GSK3beta) inhibitors for Alzheimer's disease: design, synthesis, and characterization of pyrazines. *J. Med. Chem.* **55**, 9107–9119 (2012).

157. Bhat, R. *et al.* Structural Insights and Biological Effects of Glycogen Synthase Kinase 3-specific Inhibitor AR-A014418. *J. Biol. Chem.* **278**, 45937–45945 (2003).
158. Luo, W., Rodina, A. & Chiosis, G. Heat shock protein 90: translation from cancer to Alzheimer's disease treatment? *BMC Neurosci.* **9**, S7 (2008).
159. Shonhai, A., Maier, A. G., Przyborski, J. M. & Blatch, G. L. Intracellular protozoan parasites of humans: the role of molecular chaperones in development and pathogenesis. *Protein Pept. Lett.* **18**, 143–157 (2011).
160. Solit, D. B. & Chiosis, G. Development and application of Hsp90 inhibitors. *Drug Discov. Today* **13**, 38–43 (2008).
161. Kokh, D. B., Czodrowski, P., Rippmann, F. & Wade, R. C. Perturbation Approaches for Exploring Protein Binding Site Flexibility to Predict Transient Binding Pockets. *J. Chem. Theory Comput.* **12**, 4100–4113 (2016).
162. Sharp, S. Y. *et al.* Co-Crystalization and In Vitro Biological Characterization of 5-Aryl-4-(5-Substituted-2-4-Dihydroxyphenyl)-1,2,3-Thiadiazole Hsp90 Inhibitors. *PLoS One* **7**, 5–12 (2012).
163. Sharp, S. Y. *et al.* In vitro biological characterization of a novel, synthetic diaryl pyrazole resorcinol class of heat shock protein 90 inhibitors. *Cancer Res.* **67**, 2206–2216 (2007).
164. Dymock, B. W. *et al.* Novel, potent small-molecule inhibitors of the molecular chaperone Hsp90 discovered through structure-based design. *J. Med. Chem.* **48**, 4212–4215 (2005).
165. Mollica, L. *et al.* Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations. *Sci. Rep.* **5**, 11539 (2015).
166. Schuetz, D. A. *et al.* Predicting Residence Time and Drug Unbinding Pathway through Scaled Molecular Dynamics. *J. Chem. Inf. Model.* **59**, 535–549 (2019).
167. Amaro, R. E. & Mullholland, A. J. Multiscale Methods in Drug Design Bridge Chemical and Biological Complexity in the Search for Cures. *Nat Rev Chem* **2**, (2018).
168. Votapka, L. W., Jagger, B. R., Heyneman, A. L. & Amaro, R. E. SEEKR: Simulation Enabled Estimation of Kinetic Rates, A Computational Tool to Estimate Molecular Kinetics and Its Application to Trypsin-Benzamidine Binding. *J. Phys. Chem. B* **121**, 3597–3606 (2017).
169. Jagger, B. R., Lee, C. T. & Amaro, R. E. Quantitative Ranking of Ligand Binding Kinetics with a Multiscale Milestoning Simulation Approach. *J. Phys. Chem. Lett.* **9**, 4941–4948 (2018).
170. Tiwary, P., Limongelli, V., Salvalaglio, M. & Parrinello, M. Kinetics of protein-ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E386–E391 (2015).

171. Tiwary, P., Mondal, J. & Berne, B. J. How and when does an anticancer drug leave its binding site? *Sci. Adv.* **3**, 1–8 (2017).
172. Casasnovas, R., Limongelli, V., Tiwary, P., Carloni, P. & Parrinello, M. Unbinding Kinetics of a p38 MAP Kinase Type II Inhibitor from Metadynamics Simulations. *J. Am. Chem. Soc.* **139**, 4780–4788 (2017).
173. Gobbo, D. *et al.* Investigating Drug–Target Residence Time in Kinases through Enhanced Sampling Simulations. *J. Chem. Theory Comput.* **15**, 4646–4659 (2019).
174. Lüdemann, S. K., Lounnas, V. & Wade, R. C. How do substrates enter and products exit the buried active site of cytochrome P450cam? 1. Random expulsion molecular dynamics investigation of ligand access channels and mechanisms. *J. Mol. Biol.* **303**, 797–811 (2000).
175. Kokh, D. B. *et al.* Estimation of Drug-Target Residence Times by τ-Random Acceleration Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **14**, 3859–3869 (2018).
176. Mollica, L. *et al.* Molecular Dynamics Simulations and Kinetic Measurements to Estimate and Predict Protein-Ligand Residence Times. *J. Med. Chem.* **59**, 7167–7176 (2016).
177. Bernetti, M. *et al.* Binding Residence Time through Scaled Molecular Dynamics: A Prospective Application to hDAAO Inhibitors. *J. Chem. Inf. Model.* **58**, 2255–2265 (2018).
178. Gaussian, I. Gaussian 03. (2004).
179. Van Der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–1718 (2005).
180. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43–56 (1995).
181. Hess, B., Kutzner, C., Van Der Spoel, D. & Lindahl, E. GRGMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **4**, 435–447 (2008).
182. Swope, W. C., Andersen, H. C., Berens, P. H. & Wilson, K. R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **76**, 637–649 (1982).
183. Miyamoto, S. & Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **13**, 952–962 (1992).
184. RStudio, I. RStudio-Open Source and Enterprise-Ready Professional Software for R. (2018).
185. Tenenbaum, J. B., De Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science (80-.).* **290**, 2319–2323 (2000).
186. Das, P., Moll, M., Stamati, H., Kavraki, L. E. & Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction.

Proc. Natl. Acad. Sci. U. S. A. **103**, 9885–9890 (2006).

187. Banisch, R. & Vanden-Eijnden, E. Direct generation of loop-erased transition paths in non-equilibrium reactions. *Faraday Discuss.* **195**, 443–468 (2016).
188. R: the R project for statistical computing. (2018).
189. Pan, A. C., Borhani, D. W., Dror, R. O. & Shaw, D. E. Molecular determinants of drug-receptor binding kinetics. *Drug Discov. Today* **18**, 667–673 (2013).
190. Besker, N. & Gervasio, F. L. Using Metadynamics and Path Collective Variables to Study Ligand Binding and Induced Conformational Transitions. *Methods Mol. Biol.* **819**, 501–513 (2012).
191. Saladino, G., Gauthier, L., Bianciotto, M. & Gervasio, F. L. Assessing the performance of metadynamics and path variables in predicting the binding free energies of p38 inhibitors. *J. Chem. Theory Comput.* **8**, 1165–1170 (2012).
192. Bernetti, M., Masetti, M., Recanatini, M., Amaro, R. E. & Cavalli, A. An Integrated Markov State Model and Path Metadynamics Approach To Characterize Drug Binding Processes. *J. Chem. Theory Comput.* **15**, 5689–5702 (2019).
193. Dijkstra, E. W. A note on two problems in connexion with graphs. *Numer. Math.* **1**, 269–271 (1959).
194. Izrailev, S. *et al.* Steered Molecular Dynamics. **4**, 39–65 (1999).
195. Maier, J. A. *et al.* ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB James. *J. Chem Theory Comput.* **11**, 3696–3713 (2015).
196. Pàll, S., Abraham, M. J., Kutzner, C., Hess, B. & Lindahl, E. Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS. in *Solving Software Challenges for Exascale* **8759**, 3–27 (2015).
197. Doudou, S., Burton, N. A. & Henchman, R. H. Standard free energy of binding from a one-dimensional potential of mean force. *J. Chem. Theory Comput.* **5**, 909–918 (2009).