Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Scienze Biotecnologiche e Farmaceutiche

Ciclo XXX

Settore Concorsuale: 03/D1

Settore Scientifico Disciplinare: CHIM/08

Exploring Kinetics and Drug Residence Time in Biological Systems through Molecular Simulations

Presentata da: Mattia Bernetti

Coordinatore Dottorato

Prof. Santi Mario Spampinato

Supervisore

Prof. Andrea Cavalli

Co-supervisori

Dr. Matteo Masetti
Dr. Luca Mollica

Esame finale anno 2018

“Se non ci credi veramente,
non esiste”

TABLE OF CONTENTS

ABSTRACT

Characterizing thermodynamics and kinetics of molecular systems is the ultimate goal of biophysics. In drug discovery, this information becomes essential. Understanding local and global rearrangements, how formation and disruption of biomolecular complexes occur, the molecular determinants involved and the preferred pathways followed, contribute to forming a solid ground for the development of new drugs. Quantifying specific kinetic parameters, such as off rates and the closely related residence time, is increasingly being incorporated in the drug optimization phase. Several experimental techniques established to study and quantify kinetic features. Conversely, the computational counterpart still faces severe challenges, such as accessing the time scales at which these slow events occur, while acquiring acceptable statistics.

During this PhD program, we explored current, state-of-the-art computational methods, and combination thereof, to study kinetic properties of pharmaceutically relevant biomolecules. In particular, we applied different protocols to three test systems. In the first case, we reconstructed the free energy surface of an intrinsically disordered protein and calculated interconversion rates between the differently folded states identified. In the second application, Markov State Models were employed to identify relevant states along the protein-ligand binding pathway. Using these states as a template, a putative pathway on which computing the free energy profile associated with the binding process was determined. As for the third test case, we performed unbinding simulations on a series of ligands and prioritized them according to their average computational unbinding time. The obtained ranking was subsequently confirmed by performing experimental assays.

Despite clear limitations, the picture arising from the studies was encouraging. Computer simulations emerged undoubtedly as a valuable instrument for assessing kinetic properties of biomolecular systems. Therefore, in light of the rapid advances in computer power expected from the upcoming years, their role as effective tools to assist the discovery of novel drug-like molecules is extremely promising.

LIST OF ACRONYMS

β2-AR	β2-Adrenergic Receptor
CK	Chapman-Kolmogorov
COM	Centre of Mass
CS	Chemical Shift
CV	Collective Variable
FES	Free Energy Surface
GAFF	General Amber Force Field
hDAAO	Human D-Amino Acid Oxidase
IDP	Intrinsically Disordered Proteins
ITS	Implied Timescales
KMC	Kinetic Monte Carlo
MD	Molecular Dynamics
MetaD	Metadynamics
MSD	Mean Square Deviation
NMDAR	N-Methyl-D-Aspartate Receptor
MSM	Markov State Models
NMR	Nuclear Magnetic Resonance
Path CVs	Path Collective Variables
PCA	Principal Component Analysis
PCCA	Perron Cluster Cluster Analysis
PDB	Protein Data Bank
PES	Potential Energy Surface
PT	Parallel Tempering
PT-MD	Parallel Tempering Molecular Dynamics
PTMetaD	Parallel Tempering Metadynamics
PTMetaD-WTE	Parallel Tempering Metadynamics in the Well-Tempered Ensemble
PT-WTE	Parallel Tempering in the Well-Tempered Ensemble
RC	Random Coil
RDC	Residual Dipolar Coupling
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuations
SAS	Sample and Select
Scaled MD	Scaled Molecular Dynamics
SPR	Surface Magnetic Resonance
TICA	Time-Lagged Independent Component Analysis
TPT	Transition Path Theory
WTE	Well-Tempered Ensemble

1. INTRODUCTION

1.1 Kinetics, besides thermodynamics

Macromolecules are the key characters of the cellular machinery. By taking part in a plethora of intricate networks consisting of functional pathways, they ensure the subsistence of cell homeostasis. Proteins, in particular, represent the major operative bricks on which this complex scheme in constructed. If, on the one hand, the physiological status is maintained by the regular activity of macromolecules, on the other hand, pathological conditions arise as a result of alterations in these functionalities. Intervening with the aim of resolving a pathological state implies a thorough understanding of the molecular basis underlying the specific, altered mechanisms. This, in turn, translates to the necessity of a detailed characterization of both structural and dynamical features of the biomolecules of interest. By investigating thermodynamic properties of biological systems, equilibrium populations can be determined. Thus, it is possible to identify the most relevant states in which a specific macromolecule can be found, and the relative stabilities can be quantified in terms of free energy differences. Most of the scientific work that has been carried out up to this date was focused at achieving this static, nevertheless pivotal, description.[1] However, a complete comprehension of the molecular mechanisms is obtained when integrating such information with the kinetic counterpart.[2] This includes determining the time scales at which the events occur, identifying the preferred pathways followed when transitioning from one state to another one, and recognizing the slowest, known as rate-limiting, steps comprised in the mechanism under study. However, achieving these goals is complicated by the transient nature of intermediate, elusive states that are difficult to observe both through experiments and computer simulations, thus justifying the more contained advancements in this prospect.

In the context of drug discovery, this integral picture becomes the essential ground field on which devising strategies and protocols leading to a desired pharmacological effect. Once the molecular basis underlying specific physiological and pathological conditions is well characterized, rational approaches basing on the available information can be developed. Thus, the activity of proteins, such as enzymes or membrane receptors, can be modulated to obtain a therapeutic response, or the access to non-physiological states

causing the onset of the pathology can be prevented. In the vast majority of cases, this is achieved by means of the interaction of small molecules with specific biomolecular targets. The ultimate goal then becomes the identification of novel drug-like molecules able to take part in the molecular mechanisms of interest producing a desired therapeutic effect.

1.2 On- and off-rates in drug discovery

Current drug discovery and development programs rely on experimental and computational supervision. Over the last decades, the advancements in technologies and methodologies on both sides have been tremendous. Nevertheless, the invested efforts are not balanced by the poor success rate in the identification of novel drugs.[3] Most frequently, drug-like molecules that appear extremely promising in the early phases of the drug discovery pipeline end up being discarded during the subsequent clinical trials. This is generally ascribed to insufficient or inappropriate pharmacokinetic qualities, evidences of unexpected toxicity, or *in vivo* inefficacy. In the latter case, the importance of kinetic parameters as a measure of the *in vivo* effect has gradually emerged.[4–6] Traditionally, the activity of drug-like molecules has been expressed in terms of the equilibrium dissociation constant, K_d, or by the half-maximal inhibitory concentration, IC_{50}, both often determined under closed *in vitro* conditions. However, in many cases the correspondence does not hold, particularly when the duration of the pharmacological effect is a considerable component of the *in vivo* efficacy. These findings started to shed light on the kinetic profiles of the compounds. Thus, there has been a shift of focus in order to include the association rate constant, k_{on}, and the dissociation rate constant, k_{off}, in drug discovery programs.[4] In particular, the aim has become a rational control of such properties during the drug optimization phase.

A fast binding drug, possessing a high on-rate, would be desirable when dealing with a short-living biomolecular target, as the chance of encounter is increased.[7] Moreover, the role of on-kinetics for ligand selectivity has been also outlined in recent studies.[8,9] On the dissociation side, devising drugs with smaller off-rates is the goal in the vast majority of cases. This translates to prolonged occupation of the binding site on a molecular target, resulting in extended duration of the effect. The term residence time, expressed as the inverse of k_{off}, was coined to designate this concept.[10] However, a drastic increase of side effects is observed in certain proteins when the residence time is extended. In such

circumstances, fast-off rates would be ideal.[11] Therefore, while the overall relevance of on- and off-rates has been now recognized, it is clear that best kinetic features should be identified on a case-by-case basis, depending on the specific macromolecular target of interest.

1.3 Gaining insights into biological kinetics via computer simulations

Determining thermodynamic properties of biological systems has been a major objective of structural biology since the early times. Thus, significant advances have been made over the years from both the experimental and computational point of view. As such, established methodologies exist nowadays in this respect. Drug discovery has certainly gained from these developments. On one hand, understanding mechanisms and structural features of relevant macromolecules contributed to creating a solid base for a rational approach. On the other hand, striking support for drug discovery and development programs was acquired. The recent scientific literature is clearly indicating that the effort should now be directed towards kinetics.[5,6] Therefore, integrating such information in current drug discovery protocols has become rather appealing and undoubtedly promising.

Experimental methods such as nuclear magnetic resonance (NMR) spectroscopy and surface plasmon resonance (SPR) gained considerable popularity in this respect. The former proved to be very effective in characterizing highly dynamic systems,[12,13] while the latter in measuring kinetic quantities.[14] However, from the computational standpoint, providing insights about kinetic features of biomolecular systems is still extremely challenging. A wide configurational space, including relevant transitions along slow degrees of freedom demands for both significant computational resources and effective strategies to make sampling accessible and efficient. Although techniques to enhance the exploration of the phase space can be exploited, time resolution is typically lost. Thus, methods to reconstruct information about the timescales need to be subsequently integrated.

In the work carried out during this PhD program, we took advantage of state-of-the-art computational methods and combination of these to gain insights into kinetic properties of biological systems bearing pharmaceutical relevance. We devised different protocols depending on the specific scientific problem to address and on the characteristic of the

considered biomolecules. Specifically, three strategies are shown, each one applied to a different test case.

In the first application, we reconstructed the free energy surface (FES) of an intrinsically disordered protein[15] by taking advantage of enhanced sampling methods.[16] Subsequently, we built a kinetic model basing on the determined free energy to provide kinetic information. Thus, besides identifying the differently folded states accessible, we were able to compute interconversion rates between these states.

As for the second test case, Markov State Models (MSM)[17] were combined to path collective variables[18] with the aim of determining the free energy profile associated to a protein-ligand binding process. MSM were employed in the initial stage in order to identify relevant states along the binding pathway. This information was then used as a template on which constructing a putative pathway for the process, as required by the implementation of the path collective variables.

Finally, in the last test case, we carried out unbinding simulations on a protein-ligand system. To this end, as reproducing unbinding is a non-trivial task, we exploited an enhanced sampling method, namely scaled MD.[19,20] A series of small molecules were considered and we prioritized them according to their average computational unbinding time.[21] Subsequently, the obtained ranking was compared to experimental data by carrying out experimental assays that allowed evaluating kinetic parameters, achieving a satisfactory agreement.

To provide a general understanding about the employed techniques, we first give some theoretical insights about the core concepts. This is illustrated in the next *Theory* chapter, preceded by a short historical contextualization on rate constants. The single applications are introduced and discussed separately in the following *Results* chapter. The discussion is then closed with conclusive remarks and perspectives about future directions.

2. THEORY

2.1 Kinetics of chemical reactions

The foundation of chemical kinetics can be found in the Arrhenius equation, an empirical expression formulated in 1889 by the Swedish scientist. According to this equation, the rate constant k of a chemical reaction can be denoted as:

$$k = Ae^{-E_a/RT} \tag{1}$$

where A is a constant, named the pre-exponential factor, that is characteristic for each chemical reaction, E_a is the activation energy, R is the gas constant and T is the temperature. The expression was essentially derived to describe the dependence of a rate constant on temperature. However, it inherently introduced and highlighted fundamental concepts for reaction kinetics. First of all, Arrhenius argued that, for the reaction to take place and products to be formed, the reactants need to first gain a certain, minimum amount of energy, that he termed the activation energy. Secondly, in order to determine the value of the rate, the energetic term is multiplied by a pre-exponential factor, a concept that was further elaborated by subsequent theories.

The collision theory, proposed independently by Trautz in 1916 and by Lewis in 1918, was devised to describe chemical reactions between simple species in the gas phase. For a bimolecular elementary reaction of the type:

$$A + B \rightarrow P \tag{2}$$

the velocity of formation of the product P can be calculated as:

$$v = k[A][B] \tag{3}$$

where k is the rate constant associated to the reaction, and $[A]$ and $[B]$ are the molar concentrations of the reactants. The central idea behind this theory is that, in order to generate products, reactants need to get in touch through collision, and collisions need to

be sufficiently energetic. A further complication comes from the relative orientation in which reactants need to find themselves for the collision to become effective. In this context, the rate can be expressed as:

$$k = P\sigma\sqrt{\frac{8k_BT}{\pi\mu}}N_A e^{-E_a/RT} \tag{4}$$

where P is a steric factor accounting for the orientation of the reactants, σ is the collision cross section, μ is the reduced mass of the reactants, N_A is the Avogadro number, E_a is the activation energy, and k_B, R and T are the Boltzmann constant, the gas constant and the temperature, respectively. The equation essentially summarizes the aspects introduced above, namely the steric requirement, indicated by the steric factor P, the encounter rate or collision frequency, computed through the product $\sigma\sqrt{(8k_BT/\pi\mu)}N_A$, and the minimum energy requirement, represented by the exponential incorporating the activation energy E_a. Provided that the dependence on temperature prevails in the exponential, compared to the non-exponential term representing the encounter rate, the expression has the Arrhenius form. Therefore, in this illustration, the pre-exponential factor incorporates the requirements for the encounter of the reacting species.

With the subsequent transition state theory, developed simultaneously in 1935 by Eyring and by Evans and Polanyi, the problem was discussed in terms of statistical thermodynamics. The theory introduces the idea of a transition state, indicated by $C^{\ddagger}$:

$$A + B \rightleftarrows C^{\ddagger} \longrightarrow P \tag{5}$$

The transition state is an activated complex in rapid equilibrium with the reactants and that turns into the product P by unimolecular reaction. Eyring's formulation for the rate constant associated to the forwards reaction was:

$$k = \frac{k_BT}{h}e^{-\Delta G^{\ddagger}/RT} \tag{6}$$

where h is the Plank's constant, $\Delta G^{\ddagger}$ is the free energy barrier between the reactants and the transition state, and k_B and T are the Boltzmann constant and the temperature. The equation tells that the more energy is necessary to form the transition state $C^{\ddagger}$, the smaller

the value of the exponential and, in turn, of the overall rate of reaction. We can make the expression explicit in terms of the corresponding enthalpy ($\Delta H^{\ddagger}$) and entropy ($\Delta S^{\ddagger}$) differences by means of the equality:

$$\Delta G^{\ddagger} = \Delta H^{\ddagger} - T\Delta S^{\ddagger} \tag{7}$$

Thus obtaining:

$$k = \frac{k_B T}{h} e^{-\Delta H^{\ddagger}/RT} e^{\Delta S^{\ddagger}/RT} \tag{8}$$

It is interesting to discuss the simple example of a unimolecular reaction in the gas phase. In this case, it is valid the equivalence:

$$E_a = \Delta H^{\ddagger} + RT \tag{9}$$

The above formula can be substituted in equation 8, thus obtaining:

$$k = \frac{k_B T}{h} e^{1} e^{\Delta S^{\ddagger}/RT} e^{-E_a/RT} \tag{10}$$

By isolating the exponential that expresses the dependence on the activation energy E_a, all of the remaining terms can be incorporated in the pre-exponential factor. In this view, the pre-exponential is a measure of the activation entropy associated to the reaction.

According to Eyring's equation, when the reacting chemicals accumulate enough energy, they always proceed to formation of the product. However, this is not necessarily always correct, and discrepancies with experiments arise. Therefore, to explicitly take into account this possibility, the concept of transmission coefficient was formulated. The transmission coefficient, expressed as κ, is a factor varying from 0 to 1 that indicates the fraction of reacting molecules that are in fact converted to product. Eyring's equation can be thus modified accordingly in order to include κ:

$$k = \kappa \frac{k_B T}{h} e^{-\Delta G^{\ddagger}/RT} \tag{11}$$

In this expression, the transmission coefficient represents an additional contribution to the pre-exponential factor. When κ equals 1, the expression corresponds to Eyring's formulation and energy is the only requirement for product formation. Conversely, with a value of 0, no product is obtained. What affects the value of κ is essentially the nature of the local environment in which the reaction takes place. This formulation can be attributed to Kramers, that in 1940 presented his interpretation of the transmission coefficient as a function of the solvent viscosity.[22] Kramers' theory is based on the Langevin equation that describes the motion of a body with mass *m* along a generic reaction coordinate *x* as:

$$m\frac{\partial^2 x}{\partial t^2} = -\frac{\partial U(x)}{\partial x} - \gamma m\frac{\partial x}{\partial t} + f(t) \tag{12}$$

Truncating the expression to the first term on the right side of the equivalence returns the traditional Netwon's equation of motion. Under the Langevin regime, Newton's equation is augmented by two additional forces: a friction force, measured by including the solvent viscosity γ, that dissipates the energy of the body, and a random force, *f(t)*, that acts agitating the body in a stochastic manner. Starting from the Langevin equation, Kramers carried out elegant mathematical operations to calculate the dependence of a reaction rate on γ. Without going into further details, we mention that Kramers found two solutions for limiting regimes of γ, that is very high γ, named the Kramers high friction regime, and very low γ, termed the Kramers low viscosity regime. However, he was not able to find any formulation for the rate constant that worked for the full range of γ. This point was subsequently addressed by Pollak and coworkers.[23]

Following the given illustration about rate theories, an increasing level of complexity can be observed. Starting from the more general, empiric expression conceived by Arrhenius, we arrived at presenting the more complete interpretation given by Kramers. In the context of the present thesis, where it comes to dealing with biomolecular systems plunged into liquid solvents, Kramers' theory is undoubtedly more appropriate to achieve a better description of the dynamics involved.

2.2 Exploiting computational methods to study kinetic properties of biological systems

When we talk about kinetics of biological macromolecules, we aim at describing the real dynamics of these systems. In other words, one wishes to follow their journey along different, possible states that they are able to visit and populate. A transition from an initial state to a different one might consist in a conformational change. This would be the case when folding or unfolding take place, or when a major structural rearrangement occurs due to intrinsic propensity or in response to a trigger event. Another example of transitioning between states is represented by ligand binding or unbinding. In this case, the picture is complicated by the necessity of different bodies not only to approach, but also to get in touch in specific regions of their surfaces, and potentially with specific orientations and conformations. A further source of complexity would be a conformational change that needs to take place before the final protein-ligand complex can be formed. What is common to such plethora of different scenarios is the fact that the events involved can be considered very slow in all cases. With slow, we mean that such transitions require considerably higher time scales to happen compared to local fluctuations of the atomic coordinates. In the best case, experiments succeed at describing and quantifying these events at a macroscopic scale. From a computational standpoint, the most appropriate technique to follow the dynamics of a macromolecular system over time is Molecular Dynamics (MD) simulations.[24] However, plain MD does not guarantee the access to these timescales that we are usually interested in. First of all, when performing MD, we deal with a microscopic representation of the system, and, secondly, the allowed time-resolution for the dynamics of these microscopic systems is considerably confined. This discrepancy is typically referred to as the timescale problem. Consequently, an increasing number of approaches have been specifically devised for improving sampling of the configurational space and for extracting longer timescales information while exploiting achievable computational resources.

In this section, we provide some theoretical background on methods and combination thereof that we employed to study kinetic properties of biologically relevant systems. Applications to three specific test cases are then shown in the subsequent *Results* chapter. Briefly, we combined enhanced sampling,[16] namely Metadynamics (MetaD),[25] with a bin-based kinetic model[26] to determine kinetic rates between differently folded states of an intrinsically disordered protein. As for the second test case, a Markov State Model (MSM)[17] was constructed to identify relevant states along the binding pathway of a protein-ligand system. Subsequently, such states were employed to build a guess path on which applying the path collective variables (path CVs).[18] Thus, information from MSM

was integrated with path CVs in order to reconstruct the free energy profile along the binding process, in this second example. Finally, unbinding simulations were performed for a series of ligands of a test case protein taking advantage of another enhanced sampling method, namely scaled molecular dynamics (scaled MD).[19,20]

Firstly, we give a basic theoretical overview about these techniques that we took advantage of in order to improve the sampling of the configurational space. Secondly, we discuss the methods employed to recover kinetic information from the molecular simulations.

2.2.1 Sampling the configurational space

As introduced above, due to the timescale problem, we cannot always rely on plain MD to adequately sample a configurational space involving slow events. Inspired by this necessity, the so-called enhanced sampling methods have been developed.[16] As suggested by the term, common to this class of techniques, independently from the specific strategy underlying, is an improved, still statistically correct, exploration of the configurational space. Broadly speaking, we can recognize two major subclasses. On the one side, we have methods based on reaction coordinates that describe slow degrees of freedom of the system. These reaction coordinates, referred to as collective variables (CVs), are used to guide the exploration of the phase space. Approaches such as umbrella sampling,[27] steered MD[28,29] and MetaD[25] belong to this class. On the other side, we have non CV-based techniques of different flavors, in which sampling is enhanced along all of the degrees of freedom of the system. Part of this second class are for instance tempering methods, such as parallel tempering (PT),[30] and scaled MD.[19,20]

In this work, we took advantage of both CV-based and non-CV-based enhanced sampling methods. In the following, we provide a brief description of the theory behind the techniques that we employed.

2.2.1.1 Metadynamics

In MetaD,[25] we add a history-dependent bias potential along specific reaction coordinates of our system, the CVs. These essentially describe slow degrees of freedom of the system, that allow guiding sampling towards relevant regions of the phase space.

Therefore, by taking advantage of CVs, the exploration of the highly dimensional phase space accessible to the system is reduced to a lower dimensional problem. The bias potential added through MetaD is expressed by means of the following function:

$$V_G(q,t) = \sum_{t=0,\tau,2\tau,..} We^{-(q-q(t))^2/2\sigma^2} \quad (13)$$

It has the form of a Gaussian possessing width W and height σ, which is deposited along the CV q at increasing intervals τ of the simulation time. The overall bias $V_G(q,t)$ deposited at time t is then given by the summation over the total amount of Gaussians deposited. What happens in practice is that, at time intervals τ during the simulation, the value of the CV is computed and a small Gaussian deposited on that specific point of the CV space. This Gaussian placed has now a repulsive effect, as it discourages the system to visit again that region of the CV space, thus exhorting the exploration of non-previously visited ones. Therefore, if we start our simulation with the system located inside a certain local energy minimum, the bias is going to favor first the exploration of the CV space belonging to that basin. As shown in Figure 1, we can look at the bias as if we were gradually filling the energy basin.

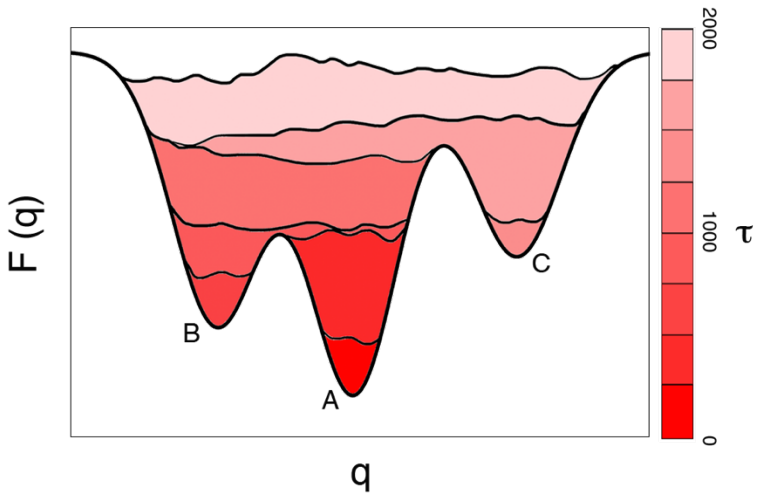

Figure 1. Pictorial representation of the MetaD method. The bias (filled curves from red to light pink) is gradually deposited at increasing time intervals τ (going from red to light pink) along a CV q. Assuming that a simulation is started from inside basin A, the gradual filling due to bias deposition allows crossing the barrier and visiting basin B. Accordingly, basin C is eventually sampled. Once all of the relevant minima have been visited, the system experiences free diffusion along the reaction coordinate. At that point, minus the total bias accumulated gives the free energy.

Suppose we are in basin A, at some point this would be filled and the system would fall inside basin B, and so on. Once all of the basins are filled, the system is able to freely diffuse in the CV space. We can then reconstruct the FES in such CV space as:

$$V_G(q, t \to \infty) = -F(q) \tag{14}$$

When employing MetaD, two major drawbacks are encountered: assessing the convergence of the simulation and choosing the CVs to bias.[31] As for the first one, it is not obvious when to stop the simulation. Once all of the basins are visited, and while the simulation keeps running, the bias keeps being deposited. This has the effect of overfilling the underlying FES and inducing 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. However, this is not straightforward to assess. In order to overcome this limitation, a variant of MetaD has been introduced, referred to as well-tempered MetaD.[32] While in standard MetaD Gaussians of constant height W are deposited over time, in well-tempered MetaD the bias is gradually reduced. As a result, the overfilling mentioned above is less pronounced. In this variant of MetaD, Gaussian height becomes a function of the simulation time, according to:

$$W(t) = W_0 e^{-V_G(q,t)/k_B \Delta T} \tag{15}$$

where W_0 is the initial Gaussian height, and $V_G(q,t)$ is the total bias deposited at time t. In the equation, ΔT represents the upper limit of the temperature range at which the CVs are sampled. In regular MetaD, the gradual addition of the potential energy bias in the CV space corresponds to sampling along the chosen CVs at increasing temperature values. Conversely, in the well-tempered variant, this behavior is confined and the mentioned overfilling of energy basins alleviated.

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 a result, the bias potential tends to smoothly converge in the long time limit. Particular care is necessary when deciding the entity of decrease in Gaussian height per time unit. W should not become too small before the basin is completely filled, otherwise not enough bias would be gathered to overcome the barrier and the system would remain stuck inside the

minimum as a result. In the current implementation, this can be controlled by setting for the simulation a specific parameter called bias factor, defined as:

$$\gamma = \frac{T+\Delta T}{T} \tag{16}$$

where ΔT is the upper limit of the temperature range to which the sampling of the CVs is confined.

Through this scheme, the potential smoothly converges in the long time limit. However, it does not fully correspond to the underlying free energy, as:

$$V_G(q) = -\frac{\Delta T}{T+\Delta T} F(q) \tag{17}$$

The second limitation is about the identification of an appropriate set of CVs. As already discussed, all of the relevant degrees of freedom of the system need to be taken into account by the CVs. If this is not the case, the simulation will not converge and pathological behaviors are likely to be observed. Identifying relevant CVs involves probably the largest initial effort, and is more a trial and error procedure. Nevertheless, it is worth noting that the effort is not completely vain, as much understanding about the system under study is typically gained at this stage. 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,[33] in which multiple replicas of the system are simulated in parallel, and a different CV is biased in each replica. Another possibility, specifically devised to manage particularly complex reaction pathways, is the use of path CVs. Conversely, the problem can be in part alleviated by coupling MetaD with non-CV-based methods, such as parallel tempering.[30] In the latter case, temperature is exploited to facilitate transitions along all of the degrees of freedom that might be neglected by the choice of CVs.

2.2.1.2 Path Collective Variables

As already widely discussed, plain MD is extremely limited when dealing with rare events. Achieving satisfactory sampling for a wide phase space, possibly comprising significant energetic barriers, would require an impressive computational effort, if

achievable at all. The risk for the system is to station within local energy minima for long periods of time, and infrequently crossing barriers in favor of different regions. Obtaining proper statistics for the estimation of an accurate FES becomes unfeasible in such a scenario. As a result, a plethora of strategies have been devised that enhance sampling and thus optimize the computational effort.[16] MetaD,[25] described in detail, is one of such methods and allows an efficient reconstruction of complex FES. However, as already discussed, a major limitation of this technique is the initial effort required to identify a valid set of CVs to guide sampling.[31] On the one hand, the chosen CVs need to take into account all of the slow degrees of freedom of the system. If this is not true, then the simulation will not converge. On the other hand, the number of CVs must be contained to avoid exceedingly slow filling times. Essentially, one needs a reasonable number of relevant CVs. This point becomes particularly challenging when it comes to protein-ligand binding, as a plethora of slow degrees of freedom can be potentially involved, ranging from solute desolvation, conformations of the ligand and rearrangements in protein residues. Therefore, with the purpose of overcoming the difficulty of managing highly dimensional phase space and reducing human intervention on the choice of the CVs, the path CVs formalism has been developed.[18]

Let us assume that we have an idea of what happens along a complex reaction, and we are able to describe it by means of a putative pathway, which is a series of frames capturing the system at intermediate states along the reaction of interest. Then we can exploit this frameset to guide sampling along the "guess path" by means of the following CVs:

$$s(X) = \frac{1}{P-1}\left(\frac{\sum_{i=1}^{P}(i-1)e^{-\lambda\left(X-X(i)\right)^2}}{\sum_{i=1}^{P}e^{-\lambda\left(X-X(i)\right)^2}}\right) \tag{18}$$

$$z(X) = -\frac{1}{\lambda}\ln\left(\sum_{i=1}^{P}e^{-\lambda\left(X-X(i)\right)^2}\right) \tag{19}$$

For a certain microscopic configuration X of the system during the MetaD simulation, the variable s can range between 0 and P, where P is the number of frames comprised in the frameset. The summations run over each frame i in the frameset, and for each of them the difference $(X\text{-}X(i))^2$ is the distance between the configuration X of the system and the one adopted in frame i. Notably, there is no restriction in the distance metric used.

However, it is common practice to use the mean square deviation (MSD), resulting in squared distance units. This is preferred over the RMSD for numerical reasons, but they are conceptually interchangeable. 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 thought as orthogonal to *s*, and expresses the distance from the putative pathway. While we advance along the putative pathway, *z* allows exploring adjacent regions of the phase space. To give a practical illustration, we can imagine the configurational space accessible to the system as confined within a cylinder, where the axis corresponds to the frameset, and *z* defines the radius. In the above functions, λ is a tunable parameter that ensures 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 has been suggested:

$$\lambda = \frac{2.3(P-1)}{\sum_{i=1}^{P-1}|X_i - X_{i+1}|} \tag{20}$$

In summary, the highly dimensional phase space is reduced to a 2-dimensional description exploiting as CVs the progression along a putative 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.

However, the non-trivial point here is that the overall framework is based on the availability of a putative pathway, that is an a priori idea of what is going on along the complex reaction. There is no general rule to obtain the required frameset, and the feasibility is strictly dependent on the information one can access. Additionally, a valid frameset responds to specific requisites. First of all, subsequent frames need to describe unidirectional progression towards the final state. No loops leading back and forth should be present. Secondly, equal spacing between subsequent frames is required. This is expressed in terms of the metric used to parameterize the guess path. 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 spacing achieved.

2.2.1.3 Parallel Tempering

As already discussed, several methods have been developed over the years to improve sampling of the complex configurational space accessible to molecular systems. While techniques based on CVs can be very efficient, they nevertheless require a considerable intervention and effort by the user. Indeed, identifying reaction coordinates that represent slow degrees of freedom of the system is far from trivial, if achievable at all. Conversely, in non CV-based techniques the bias affects the entire system, acting along all of the degrees of freedom.

PT[30] belongs to the non-CV-based class of methods. As suggested by the name, N replicas of the system are simulated in parallel at increasing temperature values T_1, T_2, T_3, … T_N. At fixed intervals during the simulation, exchanges of configurations between adjacent replicas are attempted. The probability of accepting the exchange responds to the Metropolis Monte Carlo Scheme:

$$P(i \to j) = min\{1, \Delta_{ij}^{PT}\} \tag{21}$$

with:

$$\Delta_{ij}^{PT} = \left(\frac{1}{k_B T_i} - \frac{1}{k_B T_j}\right)\left(U(x_i) - U(x_j)\right) \tag{22}$$

where $U(x)$ is the potential energy associated to configurations x_i and x_j of the system at temperatures T_i and T_j, respectively.

At the higher temperature replicas, transitions over moderately high free energy barriers are facilitated along all of the degrees of freedom of the system. At the same time, the canonical distribution is guaranteed at the lowest temperature T_1. The latter represents the reference replica, and is set to the real temperature of interest for the simulation. Therefore, the procedure has the clear advantage of including states, in the configurational space sampled at T_1, that would not be easily accessible to a traditional simulation performed at that temperature value.

According to equation 21, the acceptance probability for an exchange is strictly dependent on the overlap between the energy distributions of neighboring replicas. The acceptance ratio between adjacent replicas, expressed as the ratio between attempted and successful exchanges, is here an important parameter. By monitoring the acceptance ratio,

we assess that an adequate number of replicas have been appropriately distributed in the considered temperature range, as to guarantee sufficient overlap that allows for the desired exchanges. It is worth mentioning that this aspect is also related to the efficiency of the method. In the ideal situation, one would set the PT framework to achieve a constant acceptance ratio between any two neighboring replicas in the considered temperature range. This would ensure an optimal round trip time across temperatures, that is the time required for a cold replica to move to the highest temperature and come back, and is a measure of the efficiency of the setup.

Energy fluctuations for a molecular system increase with temperature. As such, going from T_1 to T_N, the difference in temperature between adjacent replicas can be gradually increased in order to maintain the same extent of overlap, and thus a similar acceptance ratio. Traditionally, a geometric progression of temperatures was suggested:[34]

$$T_i = T_{min}\left(\frac{T_{max}}{T_{min}}\right)^{\frac{i-1}{N-1}} \tag{23}$$

where T_i is the temperature value for each of the N replicas i, comprised between the temperature of the canonical replica, T_{min}, and the one of the replica at the highest temperature T_{max}.

It is worth noticing that the requirement of overlap between the potential energy distribution is also causing the major drawback when dealing with PT. That is, in order to cover a certain temperature range, a significant amount of replicas is typically required. This, in turn, tends to translate to high computational costs to carry out efficient PT simulations. As discussed below, this practical issue can be attenuated by taking advantage of the well-tempered ensemble.[35]

2.2.1.4 The well-tempered ensemble

As already mentioned, with MetaD we guide sampling along specific CVs. These are usually functions based on geometric criteria, such as distances, angles or the RMSD. However, we can also bias the potential energy of the system. If this is carried out through the well-tempered declination of MetaD, then the simulation is taking place in what is called the well-tempered ensemble (WTE).[35] The resulting potential energy becomes:

$$U_\gamma(x) = U(x) - \left(1 - \frac{1}{\gamma}\right)\left[U(x) - \frac{\ln N(U(x))}{k_B T}\right] \quad (24)$$

where γ is the bias factor applied, and $N(U(x))$ is the number of states with potential energy $U(x)$. It has been shown that this procedure has the effect of enhancing the fluctuations of the potential energy, while preserving the same average potential energy as in the unbiased ensemble.

Since its first formulation, the potential of this approach has been highlighted in its combination with PT. By simulating in the WTE at each replica of a certain PT scheme, the energy fluctuations at each temperature are significantly increased, to an extent determined by the bias factor applied.[36] As a consequence, the same continuous desired overlap can be achieved while placing the replicas at a farther distance between each other. Covering a certain temperature range requires, in turn, a more contained number of replicas. The overall computational effort required is thus considerably reduced and the efficiency of the method improved.

2.2.1.5 Scaled Molecular Dynamics

Another non-CV-based technique is scaled MD.[19,20] When applying this method, the potential energy surface (PES) of the system is scaled by multiplying it for a factor λ comprised between 0 and 1. When λ equals 1, then the PES corresponds to that obtained from a plain MD. As we go to lower values of lambda, approaching the 0 value, the energy profile is increasingly smoothed. Figure 2 gives a pictorial representation of the effect of λ on the PES. As a result, as easily understandable from the picture, the energy barriers between different states are lowered and thus transitions between them are facilitated. This behavior is equivalent to sampling at high temperatures. However, contrary to scaled MD, this would require shorter time steps, inevitably reducing the efficiency of the simulations.[19]

The method was initially proposed by Tsujishita et al.,[20] and was subsequently reconsidered and exploited by Sinko and coworkers.[19] Under scaled MD conditions, the canonical probability distribution for a given state of the system is modified to:

$$p^*(x) = e^{-\lambda V(x)/k_B T} \quad (25)$$

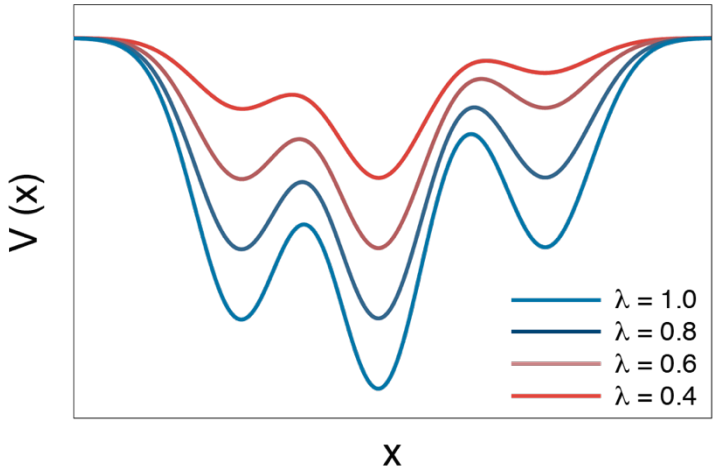

Figure 2. Effects of the scaling factor on the potential energy surface of the system. As more aggressive λ values are applied (that is, smaller), the potential energy profile is increasingly smoothed.

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

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

$$p(x) = p^*(x)^{1/\lambda} \tag{26}$$

It is worth noticing that it is possible to envisage other reweighting schemes, including for instance an energetic term. However, it was shown that the population-based strategy is more accurate, as energy terms are subjected to larger energy fluctuations that introduce larger errors.[19]

2.2.2 Recovering kinetics

2.2.2.1 Markov State Models

In computational biology, MSM construction is typically based on plain MD trajectory data.[17,37,38] There is bit of a shift of perspective compared to the traditional interpretation of computer simulations, as with MSM we analyze MD trajectories through a purely statistical approach. The clear advantage for computational biologists is the possibility of aggregating multiple, shorter plain MD simulations to achieve a description of longer timescale kinetics. Nevertheless, adequate sampling obtained via plain MD is necessary to

construct the model. Specifically, all of the states accessible to the system need to be visited with statistical significance. To this end, the model itself, as we construct it, is able to give guidance about those regions of the phase space for which more sampling is required. Notwithstanding this capability, considerable computational resources are typically required to gather sufficient data to construct a reliable MSM.

The approach represents a systematic way of decomposing the configurational space of the molecular system into a set of microstates. By counting the transitions in the microstate space according to what observed in the plain MD trajectories, one is able to calculate a transition matrix. Such matrix contains the transition probabilities between the microstates. Besides transitions probabilities, several useful properties of the system can be calculated. In particular, the slowest relaxation timescales, long-living states, that is metastable states, and also pathways and rates involved in the transition from one microstate to another, can be determined.

In the following sections, we provide some insights into the major concepts behind the construction of a MSM, namely state decomposition, count of the transitions and coarse graining.

2.2.2.1.1 State decomposition

The core concept behind the construction of a MSM is state decomposition.[37] Generally speaking, a MSM is essentially a transition probability matrix in which the probabilities of going from one state to a different one are stored. In the context of MSMs, the term microstate is preferred. A critical step is thus the definition of the microstates in which a certain molecular system can be found. In practice, this corresponds to decomposing a broad configurational space, characterized by a plethora of possible configurations of the system, into a finite, valid set of microstates that provide a suitable representation of the initial configurational space. This procedure is named state decomposition. In order to achieve such discrete subdivision, clustering is the natural choice, with the obtained different clusters representing different microstates accessible to the system. If we were able to compute the average transition time between all of the configurations sampled in a MD simulation, such average transition time would be the obvious variable on which carrying out clustering. Then, on a hypothetic journey in the microstate space, at each step the system would loose its memory about previously visited microstates. Thus, reaching the next microstate would be simply a deterministic function of the microstate on which the system currently lies. This is called the Markov property,

and is an essential characteristic that needs to be fulfilled by a MSM. This is true if no significant internal energy barriers exist within a microstate, while barriers are found when we are at abandoning it. In other words, for the Markov property to hold, transitions between members of the same state need to be faster than transitions between different microstates. However, determining average transition times between pairs of configurations is currently not attainable. Thus, the just introduced framework would break. Nevertheless, we can think about a kinetically relevant clustering procedure, achieved through geometric criteria. Translated in more practical terms, this means grouping together configurations according to geometric similarity. If the result is not too spread, that is configurations clustered together are sufficiently similar, it becomes reasonable thinking that transitions between them will be fast. This, in turn, is connected to clustering algorithm intrinsic properties and on the subdivision of the phase space. It is not in the intent of the current discussion to go into details, but we mention that, as for the latter point, one would be tempted by a finer subdivision of the phase space as a mean to increase similarity within clusters. However, it needs to be kept in mind that adequate statistics for each microstate is also required, and this is typically inversely proportional to the number of clusters.

From this illustration, the arising elements are a clustering algorithm and an appropriate distance metric on which applying it. By using the latter terminology, we intend variables that are able to capture the relevant dynamics of interest, distinguishing different configurations and identifying similar ones relatively to the process under study. In other words, variables that allow us effectively estimating geometric, and thus possibly kinetic, similarity between two structures. To give some examples, we might want to use dihedral angles to describe a small molecule possessing many degrees of freedom, while we might prefer distances between α-carbons to deal with a major conformational change in a protein. As in many cases, there is no general rule and a distance metric that fits all of the possible scenarios, thus the choice is left to the user depending on the specific scientific problem. In a nutshell, when constructing a MSM, as a first step we map all of the sampled configurations on the chosen distance metric.

For what concerns clustering, a plethora of algorithms have been reported in the literature over the years. Among these, K-medoids, K-means, K-centers, regular spatial and regular temporal have been reviewed for application in the context of MSMs.[39] By examining published papers on MSMs,[40–42] we deduced that a popular choice in the MSM community is the K-means clustering algorithm.[43,44] In K-means, data points belonging to

the same cluster present minimal pairwise distances. This is achieved by minimizing the following function:

$$J = \sum_{j=1}^{k} \sum_{i=1, x_i \in c_j}^{n} d\left(x_i - c_j\right)^2 \quad (27)$$

where k is the number of clusters, n is the amount of data points contained in the cluster, and $d(x_i\text{-}c_j)^2$ is a function estimating the distance, in units of the distance metric selected, between the data point x_i and the cluster center c_j. In subsequent iterations, the cluster centers are recalculated and data points reassigned to the closest centroid. The iteration keeps going until cluster centers do not change anymore.

One reason of K-means popularity, in the first place, is its tendency to generate microstates including a more similar number of configurations. Put differently, the algorithm creates more clusters in more sampled regions. On the one hand, the clear advantage is that this feature helps guaranteeing more reliable estimates of transition probabilities, since, as previously introduced, satisfactory statistics is required for each microstate in order to build a meaningful MSM. On the other hand, there are drawbacks when using K-means. First of all, the described behavior might exacerbate resulting in an over-division of some regions and under-division of others. Moreover, when applying K-means, the amount of clusters has to be specified by the user, thus requiring some sort of iteration in order to select an appropriate number. Finally, a cluster center is characterized by the average of the variable values from the configurations belonging to that cluster. Thus, in molecular systems, they usually do not correspond to physically meaningful states. Nevertheless, other strategies can be employed to easily identify representative structures for the generated clusters.

As a final remark, it is important to mention the possibility of reducing the number of dimensions before performing clustering. The most widespread technique for this purpose in computational biology is undoubtedly principal component analysis (PCA).[45,46] In PCA, a covariance matrix is calculated between chosen variables in order to identify those vectors on which the highest variance of the data can be projected. For instance, if we use Cartesian coordinates from a MD trajectory as input variables, the method prioritizes vectors along which the largest motion has been observed during the simulation. Another possibility, that is undoubtedly gaining increasing attention, is time-lagged independent component analysis (TICA).[42,47,48] The underlying concepts are similar to those of a PCA,

however here the focus is on the timescales. Instead of weighting vectors according to the variance, we prioritize those vectors that incorporate the slowest motions. To give an example, let us consider the MD simulation of a globular protein comprising structured domains and loops. Now, let us assume that no significant structural rearrangement took place, while the most significant fluctuations are observed in non-interesting loop regions. If we performed PCA on all of the α-carbon coordinates, the first components would likely comprise loop motions as they were the widest in terms of displacement. Conversely, if we performed a TICA, we would not likely see loop motions as the relevant component, because they are typically fast motions and thus less relevant in term of timescales. In the first place, our choice of variables could result in an extremely large number of dimension, thus we might just want to contain it. Differently, we could consider performing TICA in a pure kinetic perspective. Since its purpose is focusing on the slowest motions, our data would somehow result cleaned of the least relevant information in this sense. However, it is worth noticing that interpreting outcomes from dimension reduction techniques is not straightforward, and moving our distance metric towards a component space might result in a less intuitive microstate definition.

2.2.2.1.2 Counting the transitions

Through state decomposition we discretize a configurational space, thus allowing for a computationally manageable description of a molecular system's dynamics. Essentially, we identify clusters that correspond to possible states, referred to as microstates in this framework. In a MD trajectory, each frame corresponds to a specific configuration of the system, which can be assigned to a certain microstate among those identified. By carrying out this assignment, we can convert our conventional trajectory, that is a series of structures over time, into a discrete trajectory, that is a series of microstates visited over time. In a more general view, we are doing nothing more than assigning data points to previously defined clusters. With these discrete trajectories in our hands,[37] we can follow the journeys taken by a molecular system around the microstate space. In particular, we can make jumps of fixed size while accompanying the system on these journeys. As each of these jumps substantially corresponds to a transition from a certain microstate to a different one, we can record the number of transitions observed between each pair of microstates and store them in a matrix. The result is a count matrix C. Thus, for each point ij of this matrix, C_{ij} is the number of transitions between microstates i and j. Ideally, in the

limit of ergodic sampling, each count C_{ij} could be converted in the corresponding transition probability T_{ij} by simply maximizing the likelihood of the function:

$$T_{ij}(\tau) = \frac{C_{ij}}{\sum_k C_{ik}} \tag{28}$$

where the summation runs over all of the possible k transitions observed from state i, and τ is the jump size. With a transition probability matrix in our hands, we would be able to determine the slowest motions implied in the dynamics under study, metastable states populated by our system, and the most likely pathways leading from a certain microstate to another one. However, several issues arise in real life, such as finite sampling for the microstates, and imperfections in state decomposition, the so-called discretization error. As a result, obtaining transition probability matrices translates into more complicated routines that we will not go through in the current illustration.

Focusing back on the transition counts, a criterion for jumping over the discrete trajectories needs to be considered. As already mentioned, the advancement has a fixed size τ, so we could just keep moving forward for n steps until we reach the end of the considered discrete trajectory. In this more intuitive scheme, shown in Figure 3A, we are gathering independent counts. This would work in the limit of infinite sampling, or at least of sufficient sampling for the slowest relaxation time. Again, this is not the case in real life, and it would cause imprecise estimation of transition probability matrices. The reason is that a large fraction of the data is neglected as we jump over it, thus less statistics is obtained for each point of the count matrix.

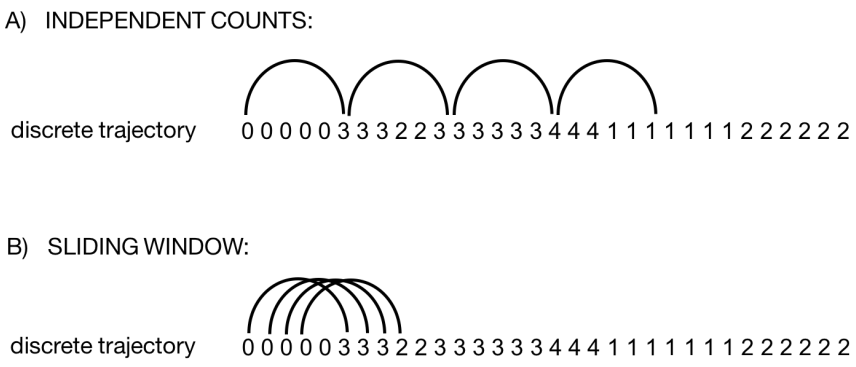

Figure 3. Schematic representation of the two methods for counting transitions over discrete trajectories. The series of numbers represents a sample sequence of microstates, that is a sample, short discrete trajectory.

Thus, in the vast majority of cases, it is common practice to exploit a sliding window scheme, as depicted in Figure 3B. Contrary to the previous strategy, in this case we consider all of the available information. However, it has been noted that this has the effect of underestimating the model uncertainty, so it needs to be taken into account when assessing it.[37]

Besides state decomposition, a second critical element when constructing a MSM is the size of the jumps. This is called the lag time of the model, and, as already shown, is expressed as τ. Roughly speaking, it coincides with a multiple of the input trajectory time step, intending the frequency with which frames have been saved and not the integration time step of the simulation. The lag time somehow defines the highest resolution of the model. In other words, we cannot expect to be able to extract information about events occurring at timescales smaller than τ. The lag time corresponds to the Markov time, which is the smallest size used to gather the transition counts that guarantees Markovian behavior. The approach that is typically followed to determine such lag time relies on monitoring how relaxation time scales, usually referred to as implied time scales, depend on it. As Markovian dynamics implies constancy of implied time scales on τ, we can assess at which value of τ the time scales reach a plateau, and then use such value to build the model. An example is given in Figure 4.

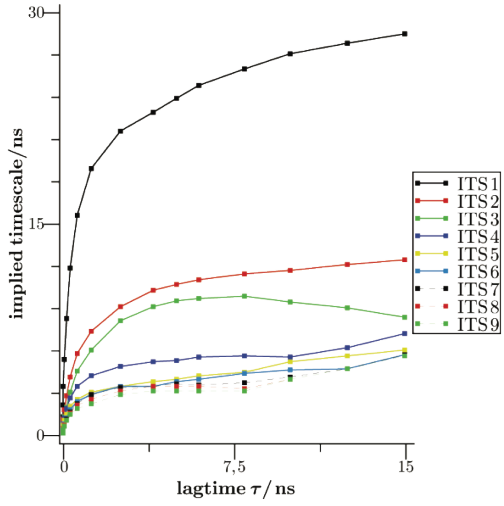

Figure 4. A typical implied timescale (ITS) plot. Relaxation timescales are calculated at increasing lag times. The Markov time is the lag time value at which the curves reach a plateau, as they become independent from τ. Adapted with permission from Ref.[39] Copyright 2017 American Chemical Society.

Given a transition probability matrix, through diagonalization we obtain the corresponding eigenvectors and eigenvalues. While the former describe specific transitions of the system, from the latter we can determine the time scales at which these motions occur. Such relaxation time scales are computed according to the function:

$$t_i = -\frac{\tau}{\ln \lambda_i} \tag{29}$$

where λ_i is an eigenvalue of the matrix underlying the MSM. According to this framework, calculating the implied time scales requires having at disposal a transition probability matrix, and this in turn means that a MSM has been constructed. Therefore, since we compute the time scales at increasing τ values in order to spot the Markov time, this translates essentially to constructing a series of MSMs at different lag times. Finally, once the eigenvalue spectrum becomes stable and independent from τ, we use that τ to build the model.

In conclusion of this general illustration, it is worth noting that another relevant aspect when constructing a model is its validation. For what concerns MSMs, a lot of effort is currently focused in this direction. Nevertheless, approaches exist already, and have been widely used, to test the robustness and thus a significant part of the validation process.[39] Among these, the Chapman-Kolmogorov test is undoubtedly the most popular.[38] For a certain transition probability matrix calculated applying a certain lag time τ, we evaluate the probability of remaining within a certain microstate A at later times $k\tau$, where k defines multiple integers of τ. The same procedure is performed on the original data set, that is the discrete trajectories, and the results are compared. This is expressed by the following function:

$$p_{MD}(A, A; k\tau) = p_{MSM}(A, A; k\tau) \tag{30}$$

and, in practice, we assess how well the equation holds. This should be performed for the achievable $k\tau$ values, according to the original dataset. With this test, we estimate the self-consistency of the constructed model, as we evaluate how well it is able to reproduce the information that has been used to parameterize it. However, the equality is not expected to hold exactly, because of statistical uncertainties due to finite sampling for the estimation of the transition probability matrix.[38] Thus, uncertainties are calculated for the probabilities

from the discrete trajectories, that is the initial MD trajectory data. Then, the result is positive if equation 30 holds within the uncertainties for the MD data. Typically, the procedure is repeated constructing several MSMs at different lag times, and showing that, at that the chosen τ, a satisfactory result for the test is achieved.

2.2.2.1.3 Coarse graining

While precious information about both thermodynamic and kinetic properties of a molecular system is engraved in a MSM, connecting back with the biology is not an easy task. An extremely wide configurational space has been discretized into a finite number of microstates. However, since the number of microstates is still very large, interpreting transition networks is challenging and typically not intuitive. Therefore, one can coalesce microstates into a reduced number of macrostates. This procedure is referred to as coarse graining the model. While, after contracting the complex picture into a set of few states, the possibility of a reliable, quantitative prediction is likely to be lost, however important indications can still be provided to direct new investigations. By coarse graining, we identify those transitions that correspond to the slowest timescales. These essentially underlie the higher barriers between the relevant metastable states of the molecular system.

Several techniques have been developed to carry out the coarse graining step. Herein, we mention Perron Cluster Cluster Analysis (PCCA)[49,50] and the more robust version PCCA+.[51,52] We remand the reader to the literature for technical insights about these methods. Herein, we give some general details about the logic behind. To facilitate the illustration, we follow Figure 5. As already said, for a given transition probability matrix, that is, for a given MSM, we can calculate eigenvectors and eigenvalues. A sample eigenvalue spectrum is shown in Figure 5C. While, as explained, the eigenvalues can be converted into timescales, the corresponding eigenvectors represent the transition that is occurring at that timescale. In reality, the components of the first eigenvector, corresponding to the first, thus largest, eigenvalue, are proportional to the equilibrium population of the states and is usually neglected. For the other eigenvectors the conversion to timescales is meaningful. In particular, from the second eigenvalue we determine the slowest timescale. The corresponding eigenvector describes the transition between the metastable states separated by the highest energy barrier. This, in Figure 5A and B, is represented by the transition between states *B* and *C*. According to the same logic, the third eigenvector expresses transitions over the second highest barrier, separating metastable states *A* and *B*, and so on.

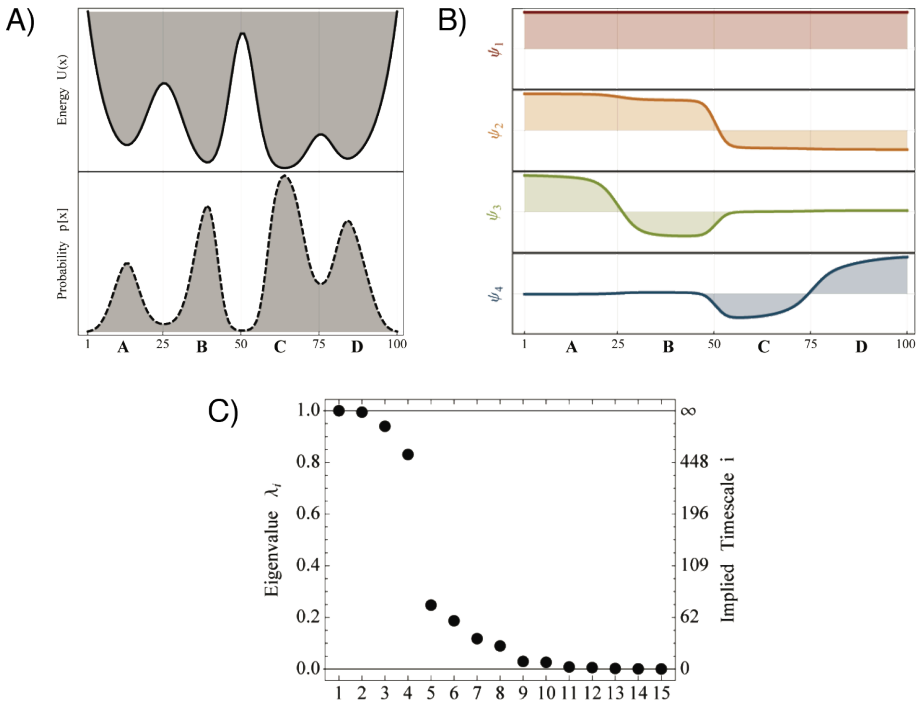

Figure 5. Coarse graining into relevant metastable states. A) A potential energy function comprising four metastable states, separated by three energy barriers. B) The four dominant eigenvectors calculated from a hypothetical transition probability matrix. C) Eigenvalue spectrum (the first fifteen are shown) and the corresponding relaxation time scales. Adapted with permission from Ref.[39] Copyright 2017 American Chemical Society.

As shown in the example, given three energy barriers, we would observe three slower timescales. This would be highlighted by the presence of a gap between the fourth and fifth values in the eigenvalue spectrum. This general framework is exploited by PCCA and several other methods to perform coarse graining. In this specific example, where three clear barriers are present, microstates would be grouped into four metastable states. However, in most cases the separation of timescales is not as clear as in the shown example. In such circumstances, the number of microstates can be interpreted as an adjustable parameter that a user can eventually comment according to the corresponding outcoming picture.[37]

2.2.2.2 Bin-based kinetic model

When it comes to sampling wide and complex configurational spaces, enhanced sampling methods are typically preferred over plain MD. Besides being hazardous, as access to all of the relevant regions is not guaranteed a priori, plain MD by definition calls for more onerous computational resources. However, when exploiting enhanced sampling

techniques, we loose completely the information about the time, and we can only discuss in terms of energetics. Given that a reconstructed FES is sufficiently accurate, and that the choice of CVs allows for an unambiguous description of the configurational space, we would be able to identify univocal metastable states and quantify the height of the barriers separating them. With this information, one should be able, at least in theory and through approximations for the pre-exponential factor, to compute kinetic rates. However, this might not be the case and other strategies need to be envisaged to extract kinetic information from biased sampling. A possible way is the construction of a kinetic model based on the free energy. Herein, we provide a general description of relevant theoretical features underlying a kinetic model based on a binning procedure of the FES.[26]

2.2.2.2.1 Kinetic Monte Carlo

As for a certain molecular system we can expect different states, a rate constant k_{ij} can be considered that describes, per time unit, the probability of escaping from a certain state *i* in favor of a certain state *j*. Transitions between different, relevant states are infrequent if compared to local fluctuations within a specific state. In other words, once it enters state *i*, the system is going to spend more time fluctuating locally before being able to leave it and reach another state *j*. Thus, transition probabilities from *i* to *j* are independent from previously visited sates. We say that the system is memoryless in this case, and the corresponding walk in the state space is called Markovian. In this picture, leaving *i* is a first-order process and:

$$p_{ij}(t) = k_{ij}e^{-k_{ij}t} \tag{31}$$

represents the probability distribution function of the first escape time towards state *j*.

For each state *i* of the system, a set of possible states *j* is accessible, and to each one of these different pathways is associated a specific kinetic rate k_{ij}. If we knew the values of each k_{ij}, we would be able to correctly describe the dynamics in the state space by means of a stochastic algorithm. One engine that is available to propagate the system along this Markov walk is Kinetic Monte Carlo (KMC).[53]

We describe herein the simplest implementation for illustrative purposes. From the exponential first-escape time distribution, we can draw an exponentially distributed random number:

$$t_{draw} = -\frac{1}{k}\ln r \tag{32}$$

where k is a rate constant associated to a certain path and r is random number comprised between 0 and 1. Basing on this, we can produce a trajectory by generating, at each step, t_{draw} for each k_{ij} associated to the accessible pathways. The lowest time t_{jmin} towards state j is going to be chosen and the overall clock of the simulation updated by adding t_{jmin}. The procedure is then repeated from the state j.

2.2.2.2.2 Bin-based procedure

With enhanced sampling methods, the bias has the effect of pushing the system towards relevant regions of the CV space. As such, transitions take place at simulation times that do not reflect the real dynamics. Therefore, transition probabilities calculated from the transitions observed in the biased trajectories cannot be used to determine real rates. However, since by means of enhanced sampling techniques we can reconstruct a FES, to which kinetic properties are intimately related, it is possible to derive free energy based models describing the kinetics of the system. Herein, we discuss a kinetic model based on a binning procedure of the FES.[26]

For a given FES described by N CVs, we can define small volumes in this CV space representing a certain state of the system, characterized by a specific value of the free energy. To this end, a binning procedure can be applied. The values spanned by each CV can be discretized in equal increments, each one of which defines one of the N sides delineating a small volume. When executing this procedure, both large and small bin sizes are not advisable. Since transition probabilities between bins are evaluated in order to construct the model, collecting sufficient statistics for each bin becomes more challenging as we reduce the size, and thus increase the number, of bins. As for large size, and thus low number, of bins, this brings to a poor description of the underlying FES, loosing resolution of the minima identified.

In this bin-based kinetic model, the transitions between pair of bins α and β are regulated by the following definition of rate:

$$k_{\alpha\beta} = k^0_{\alpha\beta} e^{-(F_\beta - F_\alpha)/2k_B T} \tag{33}$$

where $F\alpha$ and $F\beta$ are free energy associated to bins α and β, T is the temperature of interest, and k^0 is the rate corresponding to simple diffusion on a flat free energy surface, considering k^0 equal for transitions from α to β and from β to α. In this formulation, the probability of visiting bin α is proportional to the negative exponential of $F\alpha$, thus visiting lower energy states is more likely. The rates k^0 depend on the diffusion properties of the system in the reaction coordinate, that is CV, space and on the sides of the bins. In the simplest, one dimensional example:

$$k^0_{(i)(i\pm1)} = \frac{D}{ds^2} \tag{34}$$

where k^0 between neighboring bins i and $i\pm1$ is calculated from the diffusion coefficient D and the side ds of the bin in the single reaction coordinate space. In d dimensions, D becomes a matrix that stores the different values of the diffusion coefficient for dynamics in all of the possible direction of the d dimensional space. As a result, for each possible direction in bin space, k^0 is estimated from the respective bin sizes and elements of the diffusion matrix.

In order to determine the diffusion matrix, one performs several plain MD runs started at different points of the CV space. If relevant metastable states are known, the accuracy of the procedure can be improved by starting the simulations from the regions they belong to. As a good practice, representative structures can be extracted from the minima identified in the FES. The trajectories are then projected in the bin space by applying a certain lag time τ. Basing on the associated CV value, each frame of the trajectory is assigned to a specific bin. Thus, a trajectory is converted from a series of structures into a series of bins. We then make jumps of integer size τ, and record the transitions observed accordingly. This information becomes the ground on which determining the diffusion matrix. Starting from the bins visited in the MD runs and using the same τ, several KMC trajectories are started to visit the bin space with a k^0 based on an initial guess value for D. Through the KMC runs, the transition probabilities between pair of bins at lag time τ are computed as:

$$p_D(\gamma|\beta) = \frac{n(\gamma(\tau)|\beta(0))}{n(\beta)} \tag{35}$$

where $n(\beta)$ is the number of times state β was visited, and, considering to be in that state at time 0, $n(\gamma(\Delta t)|\beta(0))$ counts the number of transitions to state γ at time τ during the KMC simulation. Finally, taking advantage of these transition probabilities, we can express the likelihood of observing the same sequence of bins obtained from the MD trajectories as:

$$L(D) = log \prod_t p_D\big(\alpha(t+\tau)|\alpha(t)\big) \tag{36}$$

By maximizing equation 36 as a function of D, we eventually determine the diffusion matrix for our kinetic model. Another aspect to consider is the dependence of D on the lag time. As a typical behavior, as τ increases, D increases consequently until it converges to defined values. At that point, the dynamics in the bin space is approximately Markovian. This means that transition between states are memoryless, they only depend on the current states and not on previously visited ones. Under such circumstances, the constructed kinetic model returns a good approximation of the dynamics of the system. The τ from which the Markovian behavior is observed is called Markov time. This corresponds to the highest resolution of the kinetic model. Thus, transitions occurring at timescales lower than the Markov time cannot be reproduced. In conclusion, one maximizes the likelihood as a function of D for different values of τ, until the elements of the diffusion matrix converge to stable values.

Once D is determined, we have all of the ingredients composing the kinetic model. We can thus perform KMC trajectories in the bin space. These are started from each bin, and those on which most of the trajectories end are used to define the main kinetic basins. Finally, starting KMC runs from each kinetic basin, one can record the average time required to visit each one of the other basins and calculate the corresponding rates.

2.2.2.3 Relative residence time

The residence time of a certain ligand towards a specific biomolecular target estimates for how long that ligand is able to occupy the binding site of that target. Such quantity, determined as the inverse of the unbinding rate constant, has recently gained much interest in drug discovery.[5,6] A longer residence time for a certain drug translates to extended modulation of the activity of the target biomolecule. As such, the resulting pharmacological effect is prolonged. In the vast majority of cases, this is a desired scenario. Therefore, being able to assess kinetic profiles for drug-like molecules during the

optimization phase of a drug discovery campaign would be extremely desired. In particular, the possibility of predicting efficiently such properties for a series of compounds by means of computational methods would provide a significant improvement to the optimization pipeline. Although significant efforts are currently being focused on determination of on- and off-rates through computer simulation, assessing absolute values of these kinetic parameters still represents a major challenge. Considerable efforts are required in terms of both computational resources and user intervention for system-dependent setup. Thus, such framework does not provide appealing strategies that could be applied to several protein-ligand systems in a routinely manner.

In a typical optimization scenario, several analogues of a hit compound are considered in order to identify chemical substituents that result in an increased activity. Similarly, being able to prioritize chemicals with improved kinetic profiles would be of striking support. Recently, a methodology based on scaled MD[19,20] has been introduced that allows ranking a series of similar ligands according to their computational unbinding times.[21] Despites these are not the real residence times, they can still be employed in a relative manner to prioritize promising ligands within a series of congeneric compounds. Thus, rather than aiming at an accurate determination of the absolute values of unbinding rates, the goal is a ranking of the ligands, so as to identify those characterized by a prolonged occupation of the binding site. Relying on scaled MD, multiple unbinding events for each ligand can be sampled in reasonable times, making the procedure more suitable and appealing for real life applications. In particular, given a series of analogue compounds, several unbinding simulations can be carried out for each ligand so as to determine average computational unbinding times. After a rescaling procedure considering the λ used in the scaled MD simulations, the correlation with experimental off-rates can be evaluated by fitting to a simple linear function.

According to Eyring's equation (eq. 6) and to the subsequent developments leading to the general formulation in which the transmission coefficient κ was introduced (eq. 11), we can express the unbinding rate constant for a ligand from the binding site of a specific molecular target as:

$$k_{off} = \kappa \frac{k_B T}{h} e^{-\Delta G_{off}^{\ddagger}/RT} \tag{37}$$

where $\Delta G_{off}^{\ddagger}$ is the free energy difference between the ligand-target complex and the transition state leading to the unbound state. By incorporating all of the non-exponential contributions in a general pre-exponential factor A, similarly to Arrhenius' equation, we can simplify the notation of equation 37 to:

$$k_{off} = Ae^{-\Delta G_{off}^{\ddagger}/RT} \tag{38}$$

For two ligands binding to a certain biomolecular target under similar conditions, we can assume a similar A. Therefore, we can express the ratio the ratio between their k_{off} as:

$$\frac{k_{off,1}}{k_{off,2}} = \frac{A_1 e^{-\Delta G_{off,1}^{\ddagger}/RT}}{A_2 e^{-\Delta G_{off,2}^{\ddagger}/RT}} \approx \frac{Ae^{-\Delta G_{off,1}^{\ddagger}/RT}}{Ae^{-\Delta G_{off,2}^{\ddagger}/RT}} = \frac{e^{-\Delta G_{off,1}^{\ddagger}/RT}}{e^{-\Delta G_{off,2}^{\ddagger}/RT}} = e^{-\Delta\Delta G_{off,1,2}^{\ddagger}/RT} \tag{39}$$

According to the previously introduced equation 7, we can make equation 38 explicit in terms of enthalpic and entropic contributions:

$$k_{off} = Ae^{-\Delta G_{off}^{\ddagger}/RT} = Ae^{-\left(\Delta H_{off}^{\ddagger} - T\Delta S_{off}^{\ddagger}\right)/RT} \tag{40}$$

Therefore, by applying the scaled MD λ factor to equation 38 and considering the explicit formulation introduced in the above expression, we can write:

$$k_{off,\lambda} = Ae^{-\left(\lambda\Delta H_{off}^{\ddagger} - T\Delta S_{off}^{\ddagger}(\lambda)\right)/RT} \tag{41}$$

where $\Delta S_{off}^{\ddagger}(\lambda)$ indicates the unknown effect of λ on the entropic term. Assuming equal entropic contributions for the two considered ligands, since we manage analogues from the same scaffold that are unbinding from the same protein under the same conditions, we can cancel out the entropic term contained in the ratio introduced by equation 39. Under this assumption, and considering the form of equation 39, λ can be collected and the equation expressed as:

$$\frac{k_{off,\lambda,1}}{k_{off,\lambda,2}} = e^{-\lambda\left(\Delta\Delta G_{off,1,2}^{\ddagger}/RT\right)} = e^{-\left(\Delta\Delta G_{off,1,2}^{\ddagger}/RT\right)^{\lambda}} = \left(\frac{k_{off,1}}{k_{off,2}}\right)^{\lambda} \tag{42}$$

Being the residence time t_r the inverse of the off-rate, the above equivalence can be rewritten as:

$$\frac{t_{r,\lambda,2}}{t_{r,\lambda,1}} = \left(\frac{k_{off,1}}{k_{off,2}}\right)^{\lambda} \tag{43}$$

This relationship between unscaled and scaled parameters can be used to assess the correlation between real, experimental values and those computed at a certain λ. In this way, the ability of the method to correctly rank ligands in terms of the unbinding kinetics profile, as predicted by means of scaled MD simulations, can be evaluated.

3. APPLICATIONS

3.1 TEST CASE 1: N_{TAIL}

3.1.1 Introduction

In the last decade, it has become clear that a significant proportion of eukaryotic proteins are unstructured under physiological conditions.[54,55] The lack of a stable secondary and tertiary structure is the peculiar feature of this class, hence the term Intrinsically Disordered Proteins (IDPs). This intrinsic disorder may involve the overall amino acid sequence or just be limited to specific domains. Notably, it is responsible for the highly dynamic nature of these proteins. Indeed, rather than possessing a well-defined, most probable state at equilibrium, IDPs typically exist as a dynamic conformational ensemble of states, among which they can easily interconvert. This feature accounts for the peculiar ability of IDPs to interact with multiple targets and thus to take on many physiological roles related to, for example, signaling or regulation.[56,57] Likewise, dysfunctions of these proteins can lead to several pathological conditions, such as neurodegenerative disorders and cancer.[58,59] To gain a comprehensive picture of their functionality and implication in diseases, it is thus essential to produce a detailed characterization of the equilibrium properties of free IDPs in solution, in terms of both thermodynamics and kinetics.

Exploration of this part of the proteome is still in its infancy. Effective experimental tools for studying IDPs include NMR spectroscopy,[60] small angle X-ray scattering (SAXS),[61] and fluorescence Förster resonance energy transfer (FRET),[62] which are well-suited to dealing with the structural dynamics of proteins. However, characterizing the equilibrium properties of IDPs remains challenging due to their highly dynamic nature, as well as the heterogeneity of the above techniques in terms of space and time resolution. From a structural standpoint, NMR spectroscopy has been particularly successful in characterizing the ensemble of conformations that collectively describe these peptides. A two-step procedure called “sample and select” (SAS) is typically the most common process for identifying a representative ensemble of structures sampled at equilibrium. First, a pool of configurations is generated, usually by exploiting conformer libraries.

Then, through an iterative process, the ensemble is filtered by satisfying the best agreement between experimental and back-calculated NMR observables, such as scalar couplings (J), chemical shifts (CSs), and residual dipolar couplings (RDCs). Notwithstanding the power and elegance of this approach, SAS does not provide any kinetic information on the implied timescales regulating the interconversion between conformations. But this information is instrumental to understand the folding of IDPs and the recognition mechanisms with molecular partners. The only source of dynamic and time-resolved information by NMR is the study of spin dynamics: a method for IDPs was recently proposed,[63] based on multiple temperature measurements, to discriminate between motions on different timescales. However, even within the framework of this new family of experimental techniques, the nature of ^{15}N spin relaxation only allows the identification of motions in the timescale of tens of nanoseconds, i.e. within times that are shorter than the ones governing the interconversion between different local structures present along a disordered chain (e.g. considering the folding of a helix occurring within the μs timescale).

In this context, all-atom MD is emerging as a key tool,[24,64–67] as it provides an ensemble of conformations in equilibrium conditions. Additionally, MD can describe the kinetics of the observed events,[4] allowing a comprehensive picture of IDPs at a fully atomistic level. However, when applying MD to study slow processes such as folding and unfolding, extensive simulations (in the microsecond-to-millisecond timescale) are typically required, even for relatively short amino acid sequences.[68] MD-based enhanced sampling methods are particularly promising in this regard, as they allow for an efficient exploration of the configurational space, while preserving the Boltzmann distribution of states in the given statistical ensemble.[16] In these methods, the sampling is accelerated by either exploiting high temperature replicas, as in replica exchange (or Parallel Tempering MD, PT-MD), or through bias potentials or forces acting on selected degrees of freedom, which are known to describe the event one wishes to accelerate (reaction coordinates or collective variables, CVs). MetaD[25] is one such CV-based methods that can ultimately be combined with PT-MD[69] to further improve sampling effectiveness (PTMetaD).[70] The ever-increasing number of computational studies focusing on IDPs[71,72] demonstrates the strong need for robust techniques to simulate these systems. Such techniques range from force field optimization to the use of implicit solvent models. Interestingly, NMR can be directly combined with MD, leading to ensemble-restrained MD simulations.[73] Within these approaches, chemical shifts are implemented as additional terms of the molecular mechanics force field.[74] It has been shown that using replica-averaged restraints improves

the description of the protein dynamics around the native structure.[75] Conversely, the so-called NMR-guided MetaD incorporates experimental information in the form of a purposely designed CV. It has been successfully used to characterize the complex free-energy landscape of the Aβ1-40 peptide.[76]

The C-terminal domain (N_{TAIL}) of the nucleoprotein of the Sendai virus, a member of the Paramyxoviridae family, is intrinsically disordered.[77] Experimental studies on N_{TAIL} and similar peptides from the same family have revealed a high helical propensity in the central region of the sequence, giving rise to differently folded states that are accessible to this class of domains (Figure 6).[78]

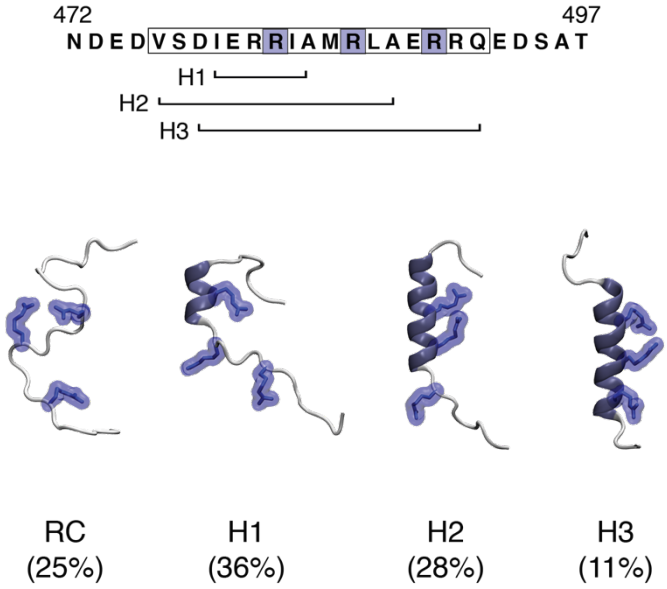

Figure 6. N_{TAIL} proposed conformational equilibrium in solution. The central region of the sequence (top, residues 476-492) encodes for differently folded states (bottom) that can be visited by the peptide. The three arginines (R482, R486, R490, in ice blue in the top sequence), through which the interaction with the molecular target PX is established, are highlighted as ice blue surface in the structures.

Herein, we investigate the ability of state-of-the-art computational methods to tackle both the thermodynamic and kinetic aspects of IDPs. In particular, we first used the enhanced sampling method PTMetaD in the WTE (PTMetaD-WTE, for clarity hereafter in this chapter simply referred to as MetaD)[35,36] to explore the conformational space of N_{TAIL} and to characterize its free-energy landscape without any bias towards experimental observables. The reliability of calculations was then validated by comparing the back-calculated CS with previously collected NMR data. Finally, a bin-based kinetic model was built to describe the interconversion between states populated by the peptide, and to

calculate the corresponding rates. Although there were marginal discrepancies with structural observables, our approach succeeded in identifying all of the relevant conformational states of N_{TAIL}, and can be useful in characterizing interconversion rates that are otherwise elusive to direct experimental investigation. In addition, the rather good agreement between theoretical and experimental NMR observables provides, to the best of our knowledge, the first quantitative assessment of MetaD to fully explore the FES of a disordered polypeptide.

3.1.2 Methods

3.1.2.1 Simulation setup

The initial system for our simulations was prepared from conformation H1 of the peptide (sequence: NDEDVSDIERRIAMRLAERRQEDSAT, top in Figure 6), determined via NMR in a previous work. Termini were capped by adding the acetyl (ACE) and N-methyl amide (NME) groups at the N- and C-terminus respectively. Solvation was accomplished by adding 15,622 water molecules in a cubic box having 78 Å edge size. Na^+ counter ions were subsequently added to neutralize the system. A last generation force-field was used to treat the peptide. In particular, the Amber ff99SB*-ILDN,[79] resulting from the original ff99SB[80] corrected with the "ILDN" side-chain torsion parameters and the helix-coil transition balance optimizations,[81] was adopted. While other computationally more expensive simulative setups could be envisioned (e.g. Amber ff03w[82] combined with TIP4P/2005[83] or Amber ff99SB-ILDN combined with TIP4P-D),[84] Amber ff99SB*- ILDN was recently used in several studies together with cheap water models, and worked well for a large variety of proteins including IDPs.[64,79,85–87] Accordingly, the TIP3P[88] model was applied for the water molecules, and the ions were described as indicated in a recent reparameterization carried out by Joung and Cheatham.[89]

After an initial minimization consisting of 5,000 steps of steepest descent, the system was equilibrated in two stages. First, we performed a 500 ps long simulation at 298 K in the NVT ensemble using the velocity-rescaling thermostat,[90] with a 0.1 ps time constant for coupling. This was followed by 500 ps in the NPT ensemble using the Parrinello-Rahman barostat,[91] with a 2.0 ps time constant for coupling. Long-range electrostatics were treated via the particle mesh Ewald method.[92] A cut-off of 12 Å was used for short-ranged non-bonded interactions. All simulations were performed applying a 2 fs time step.

For the PT, the 298-400 K temperature range was covered with 8 replicas, geometrically distributed[34] according to eq. 20, as discussed in the *Theory* chapter. Specifically, the obtained values were: 298.00, 310.80, 324.15, 338.07, 352.60, 367.73, 383.53, 400.00 K. The system equilibrated at 298 K was heated up to 400 K in 7 subsequent steps in order to obtain the starting configurations for each replica. All simulations were performed with the 4.6.7 version of the GROMACS MD engine,[93] patched with the PLUMED 2.1 software.[94]

To run in the WTE, a preliminary 5 ns long PTMetaD run was carried out, using the potential energy as the only CV. Gaussians with a height W of 2.5 kj/mol and width σ of 500 kj/mol were added every 250 steps. A bias factor of 50 was applied to ensure a sufficient overlap between the energy distribution of neighbouring replicas, and exchanges were attempted every 100 steps. The bias gathered in this preliminary run was then kept fixed during the subsequent PTMetaD production run, thus allowing simulating in the WTE. The overlap from standard PT is compared to the one achieved through combination of PT and WTE (PT-WTE) in Figure 7; to appreciate the improvement obtained, note that the same number of replicas was employed in the same temperature range.

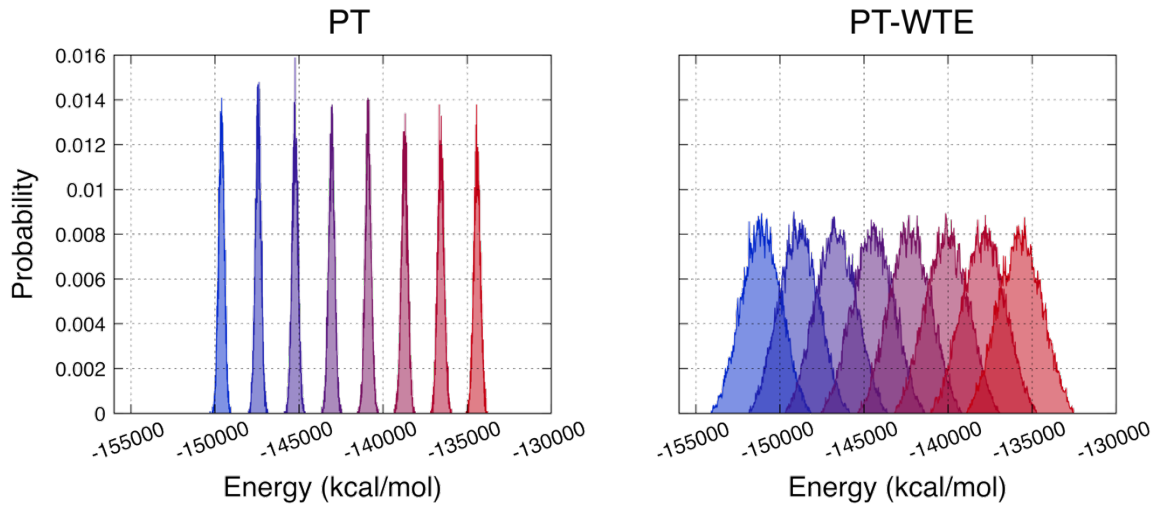

Figure 7. Energy distribution of replicas at increasing (blue to red) reference temperatures. Standard PT (left panel) is compared to PT-WTE (right panel) while employing the same number of replicas. No overlap between adjacent replicas was observed in the former case. Thus, a higher number of replicas would be required to cover the considered temperature range. The overlap was increased in PT-WTE, allowing the use of a reduced number of replicas to cover the same range.

For the production phase, the alpha helical content (α_{cont})[95] and the radius of gyration (R_{gyr})[96] were chosen, respectively defined as:

$$\alpha_{\text{cont}} = \sum_{\alpha} n[\text{RMSD}(\{R_i\}_{i\in\Omega_\alpha}, \{R_0\})] \quad (44)$$

$$R_{\text{gyr}} = \sqrt{\frac{\sum_i^n m_i |r_i - r_{\text{COM}}|^2}{\sum_i^n m_i}} \quad (45)$$

The first CV generates a set of all the possible portions comprising six contiguous residues in the amino acid sequence of a peptide. The RMSD (root mean squared deviation) between each element of this set and an idealized six residues alpha helix is then calculated. These RMSDs are finally summed to estimate the overall alpha helical content of a certain conformation of the peptide under study. The second CV is based on the distance r_i between each point comprised in an ensemble possessing center of mass r_{COM}. As points are more spread in the Cartesian space, R_{gyr} returns higher values, while it returns lower values if they are closer. Herein, the α-carbons in the peptide were considered. Thus, lower values of the CV correspond to more compact conformations of the peptide, while higher values correspond to more elongated ones. Gaussian with nominal height 0.4184 kj/mol and σ of 0.15 for α_{cont} and 0.025 for R_{gyr} were deposited every 500 steps with a bias factor of 15. Exchanges between neighbouring replicas were attempted with the same frequency as the preliminary step, resulting in an average exchange probability of about 10% in the production phase. Each replica was simulated for about 400 ns, resulting in a total simulation time of about 3.2 μs.

3.1.2.2 NMR data prediction

As reported in the past in a work published by some of the authors of the present publication,[97,98] experimental chemical shifts (δ_{exp}) have been extracted and tabulated from the original paper and put in comparison with chemical shifts predicted by means of an iterative use of an existing software. In this case, chemical shifts (specifically Cα, Cβ, C', N, HN, Hα) have been predicted (δ_{pred}) using SPARTA+,[99] a software based on artificial neural networking, applied to single structures extracted from the trajectories of each simulation every 0.1 ns, with chemical shifts calculated for each of them and then reweighted as follows:

$$\delta_{\text{pred}} = \sum_{i=1}^{N} w_{x,i}\, \delta_{\text{pred},i} \quad (46)$$

where $w_{x,i}$ represents the weight w attributed to the single structure i according to method x, and $\delta_{pred,i}$ is the chemical shift predicted for the structure i. The root mean square deviation (RMSD) between δ_{pred} and δ_{exp} for each nucleus has been calculated over all the residues j (excluding the N and C termini ones) according to the formula:

$$\mathrm{RMSD} = \frac{1}{N}\sqrt{\sum_{j=2}^{25}\left(\delta_{\mathrm{pred},j} - \delta_{\mathrm{exp},j}\right)} \quad (47)$$

A total number of 80055 structures have been extracted from the original trajectory, and four different methods have been adopted to assign weights:

a. a clustering of the structures based on the simple linkage method, i.e. a structure has been added to a cluster when its distance to any element of the cluster computed on the protein backbone between residues 473 and 496 was less than 1 Å; a total of 4335 clusters have been obtained, whereas only 594 clusters containing more than 3 members have been considered for the final chemical shifts and residual dipolar couplings calculations and statistics;
b. a binning of the FES obtained from MetaD simulations has been adopted that led to 2912 Boltzmann weighted representative structures of N_{TAIL};
c. a "sample and select" procedure has been used to select an ensemble of 7858 structures that optimize the agreement between the predicted and the experimental observables, i.e. the set that best minimizes the following RMSD-based figure of merit, namely *w*RMSD = RMSDuα + RMSDuv + RMSDuβ where RMSDCα, RMSDC', RMSDCβ, are respectively the RMSDs between experimental and computed values of Cα, C' and Cβ chemical shifts, with the last one reduced in weight due to its relative insensitivity to secondary structure variations. The limit value of wRMSD that led to the ensemble was chosen to be 2.75 ppm.
d. a sampling method based on the bin-based kinetic model, with a population attributed to the N_{TAIL} microstates after numerically solving the equations that describe the interconversion between them.

3.1.2.3 Kinetic model setup

In our binning scheme, we first explored the possibility of exploiting the free energy landscape projected onto the 2 CVs used to bias our MetaD simulation. We thus divided

the FES obtained into different combinations of N x M grids, where N and M are the number of bins along α_{cont} and R_{gyr}, respectively. However, even using finer grids (e.g. 70 x 40) this approach resulted in significantly dissimilar conformations being grouped in the same microstate. Same α_{cont} values can be shared by conformations in which the helicity is positioned in different, diametrically opposite regions along the peptide sequence (see sample structures belonging to a same energy minimum in Figure 10 of the following *Results and discussion* section). This observation led us to identify $\Delta\alpha$ as a promising, effective CV able to discriminate the location of the α-helices in the structure:

$$\Delta\alpha = \left(\alpha_{\mathrm{cont},1-13} - \alpha_{\mathrm{cont},14-26}\right) \tag{48}$$

where indices 1-13 and 14-26 refer to the first and second half of N_{TAIL} sequence respectively. Positive values of the variable indicate the helicity prevails in the first half of the peptide, while negative values in the second one.

Thus, relying on the reweighting scheme developed in previous studies, we projected the FES in the two-dimensional space of the α_{cont} and $\Delta\alpha$ CVs (Figure 13 in the *Results and discussion* section), and carried out the binning procedure introduced above. We divided our FES applying different combinations of N x M. The best compromise between number of bins (which influences microstate populations) and conformational consistency within the microstates was achieved applying a 40 x 8 grid along the CVs, α_{cont} and $\Delta\alpha$ respectively.

To determine D in our CV space, we carried out 4 plain MD simulations, 100 ns long each. The initial structures were chosen among medoids identified via the cluster analysis, and were selected to be as much distant as possible along the α_{cont} CV. For each trajectory, we computed D at increasing values of Δt. The diagonal elements of D, named *D11* and *D22* respectively, converged at Δt = 8 ns (Figure 8A). As for the off-diagonal terms, we deducted from our calculations that they need to be 2 orders of magnitude lower than the diagonal terms. For calculation of the kinetic rates, the average value of D obtained at Δt = 8 ns in the 4 simulations was applied. We also assessed the dependence of D on trajectory length at fixed Δt. One out of the 4 trajectories was extended to 200 ns, and D elements computed at increasing trajectory length, namely 25, 50, 100, 150 and 200 ns (Figure 8B).

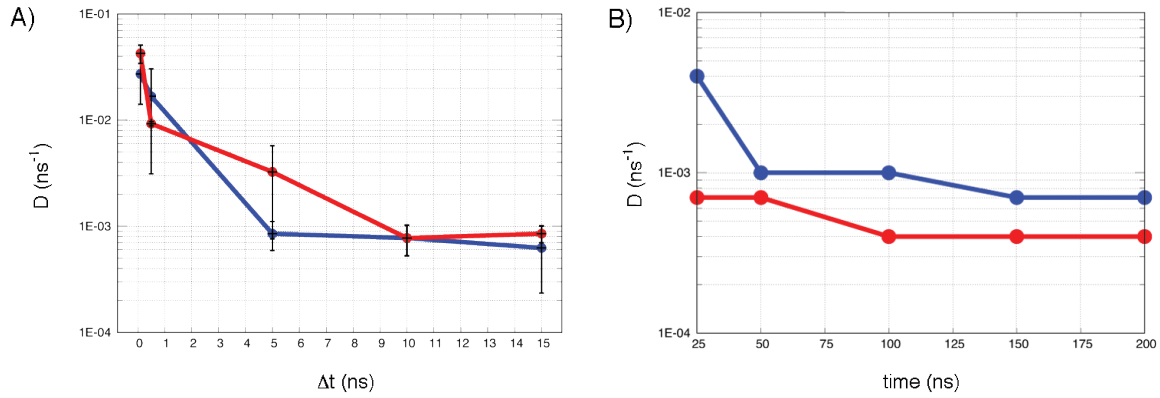

Figure 8. A) Convergence of the diffusion matrix *D*. The diagonal elements *D11* (blue) and *D22* (red) are reported as a function of the time lag (*Δt*). Error bars represent the standard deviation obtained after averaging over the 4 plain MD trajectories. B) Dependence of diagonal elements *D11* (blue) and *D22* (red) on simulation time. Y axis are shown in log scale in both panels.

While no clear-cut separation among the times scales could be identified in the eigenvalue spectrum obtained from the transition matrix computed applying the determined *D*, a noticeable slope change was at approximately the fourth or fifth eigenvector. Thus, a series of kinetic models were built including an increasing number of states, ranging from 3 to 6. In retrospect, we decided to employ the kinetic model including six states, since six was the minimum number required to observe the four experimentally derived conformations in individually distinct states. We computed the kinetic rates by performing 100 KMC simulations between each pair of the six identified states.

3.1.2.4 Calculation of equilibrium populations

Having a set of kinetic constants of formation and disruption of all the microstates identified by our kinetic model is in principle possible to compute the theoretical equilibrium concentration of each species by means of a set of differential equations that describe the rate of formation of each species present in the system. Every interconversion can be seen as a first order kinetics process that allows a transformation of one microstate into all the others and vice versa. Hence, the rate of formation/disappearance of *A* (as well as the one of the other states) can be described by the following differential equation that takes into account the coexistence of several interconversion processes:

$$\frac{d[A]}{dt} = (-k_{BA}[A] + k_{AB}[B]) + (-k_{CA}[A] + k_{AC}[C]) + (-k_{DA}[A] + k_{AD}[D]) + (-k_{EA}[A] + k_{AE}[E]) + (-k_{FA}[A] + k_{FA}[F]) = -\sum_i k_{iA}[A] + \sum_i k_{Ai}[i] \quad (49)$$

where [A], [B], [C], [D], [E], [F] indicate the concentration of microstates, k_{iA} (with i = B, C, D, E, F) are the rates of interconversion of any state i into A and k_{Ai} are the rates of interconversion of A into any other state i. In this way, the kinetics of the system can be fully described by six differential equations in the same form of equation 49, and they have been numerically solved with the help of the program for symbolic calculations SAGE[100] imposing the boundary condition:

$$[A] + [B] + [C] + [D] + [E] + [F] = \text{constant} \tag{50}$$

for respecting the physical condition of the conservation of mass during the interconversion time, and approximating the system of differential equations putting to zero the k_{iA} and k_{Ai} that correspond to interconversion times longer than 300 ns.

3.1.3 Results and discussion

We designed our experimental studies to achieve a thorough understanding of the conformational dynamics adopted by free N_{TAIL} in solution. In previous works, plain MD has been applied to sample the conformational space of proteins, and average spectroscopic observables have been predicted in agreement with experiments, thus certifying the reliability of the simulations.[98,101] Similar procedures have been presented in the context of enhanced sampling and employed to study structured proteins.[97,102] However, to the best of our knowledge, such framework has not been explored for structurally heterogeneous systems such as IDPs. Our study on the IDP N_{TAIL} revealed an ensemble comprising a random coil (RC) in equilibrium with differently folded states, involving residues ranging from 476 to 492.[78] In particular, conformation H1, in which five consecutive residues from the first half of the peptide form a small helix, has been reported as predominant by NMR experiments. We tackled this heterogeneous conformational space by exploiting MetaD. Our sampling strategy used two general purpose CVs, namely the radius of gyration (R_{gyr})[96] and the α-helical content (α_{cont}).[95] To improve the sampling effectiveness along the orthogonal degrees of freedom not explicitly included in the CVs, we used eight replicas spanning a temperature range of more than 100 K, together with enhanced potential energy fluctuations (WTE framework, see Figure 7 in

the *Methods* section). Notably, a similar setup was recently used by Han et al. to study the free-energy landscape of the closely related measles virus (MeV) N_{TAIL}.[103] Our simulations reached a satisfactory convergence of sampling after 2.6 μs (325 ns per replica, Figure 9).

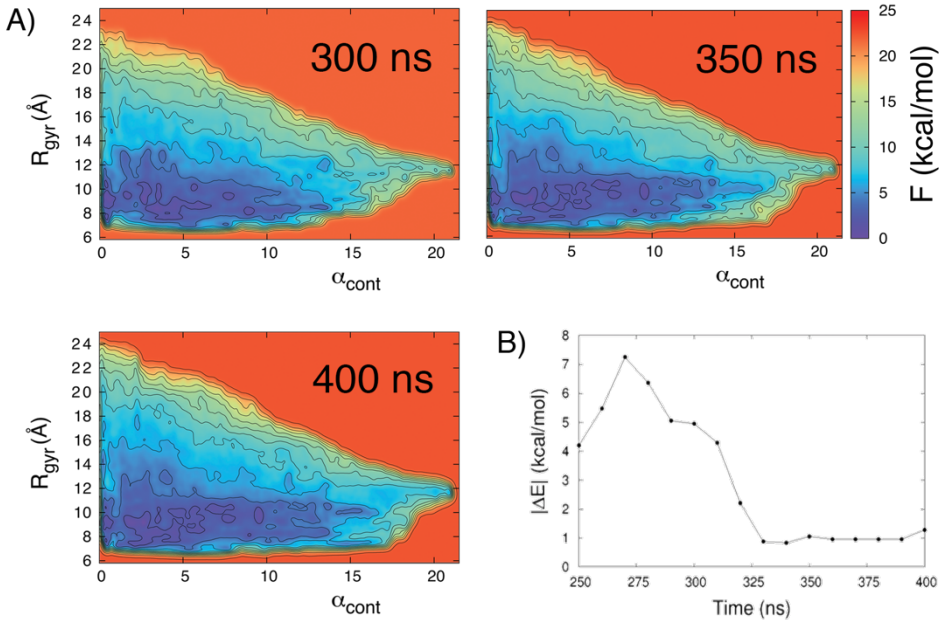

Figure 9. Convergence of the MetaD simulation. A) Free energy surface as a function of the simulation time. B) Unsigned free energy difference between the deepest minima, namely the ones located at α_{cont} = 1.5 and 9.5, as a function of the simulation time.

The result was a relatively flat and rough free-energy surface (FES), displaying multiple shallow energy minima (Figure 10). In the obtained energy landscape, the basins are mostly distributed along the α_{cont} CV, indicating the presence of a plethora of conformations possessing different and increasing secondary α-helix content. Structural characterization of these minima, achieved with a k-medoids clustering algorithm,[104] revealed differently folded states represented by α-helices mostly distributed in the peptide's central region. Notably, our simulation strategy reproduced all the relevant conformations previously identified by NMR experiments. In the free-energy landscape obtained, the relevant minima, mostly iso-energetic, are found at increasing amounts of helical content ($\alpha_{cont} \approx$ 1, 3, 6, and 9) and reasonably small values of radius of gyration, always lower than 14 Å. The random-coil conformations of N_{TAIL} were characterized by the lowest α_{cont} values ($\alpha_{cont} \approx 1$ and $R_{gyr} < 12$ Å).

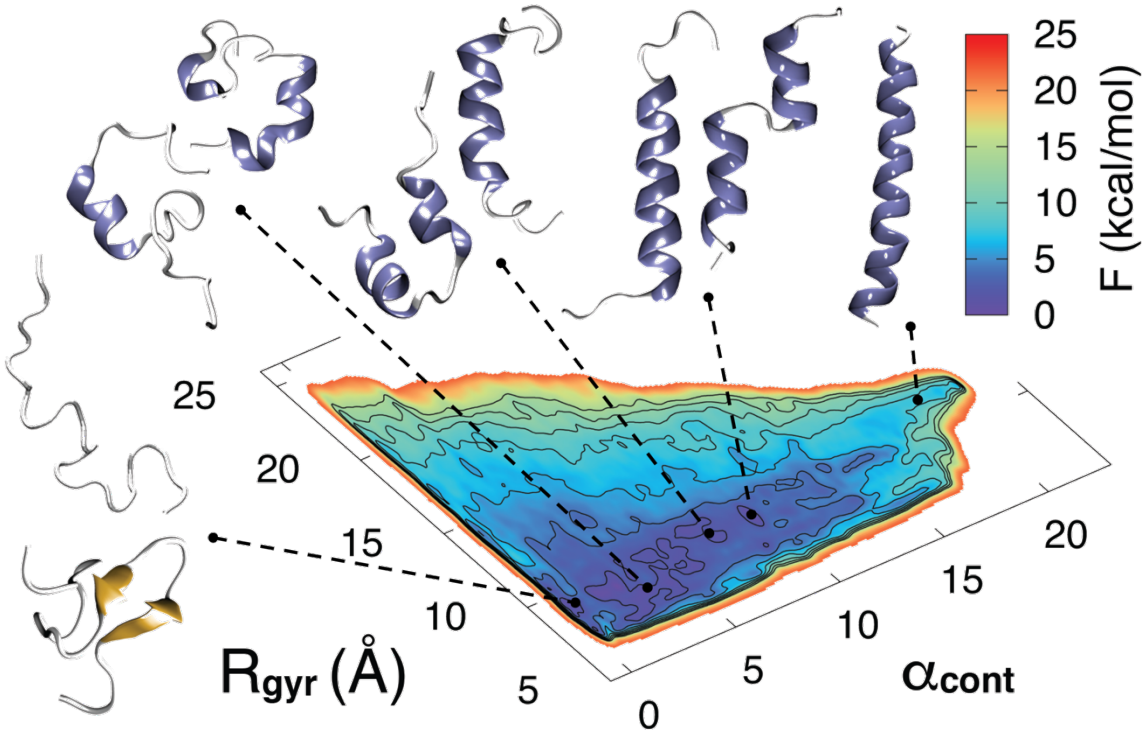

Figure 10. Free-energy surface of NTAIL obtained from the PTMetaD-WTE simulations. Contour lines are reported with intervals of 2 kcal/mol. Representative structures from each local minimum are also shown.

Notably, in this region of the CV space, we could also identify β-hairpin motifs. This reflects the ability of our MetaD setup to efficiently identify conformations bearing structural elements that were not explicitly accounted for by the CVs used. Interestingly, the most extended helix identified by NMR measurements, spanning 15 residues (conformation H3, 478-492), was sampled at approximately $\alpha_{cont} \approx 9$ and $R_{gyr} \approx 12$ Å together with differently folded intermediates. Conformation H2, comprising 13 residues (476-488), was found at $\alpha_{cont} \approx 6$ and $R_{gyr} \approx 12$ Å, in a minimum separated from the previous one, and from the one including the least folded conformation H1 ($\alpha_{cont} \approx 3$ and $R_{gyr} \approx 8$ Å). Moreover, a fully folded state ($\alpha_{cont} > 18$) was observed in a region of higher free energy ($\Delta F > 6$ kcal mol-1).

The conformational ensemble reconstructed by MetaD simulations was validated by assessing its ability to reproduce observed NMR chemical shifts. As demonstrated in previous works, single structures cannot provide accurate atomic level reproductions of polypeptide chain experimental data. However, including dynamic information dramatically improves these reproductions, encoding into single experimental observables the heterogeneity of structures due to thermal motions.

Applying the same type of approach previously adopted for structured proteins, we calculated the expected ^{13}C, ^{15}N, and ^{1}H backbone chemical shifts for each structure of the ensemble using the SPARTA+ program.[99] For all nuclei, we evaluated the RMSD between the chemical shifts estimated from the models and the experimental ones. From our calculations, it is clear that MetaD can reproduce experimental data rather well (Table 1), i.e. with an RMSD comparable with (or even lower than) those that can be calculated for structured proteins.[97,98,101] This represents a clear and quantitative indication of the ability of MetaD simulations to provide an efficient and reliable sampling of the FES and of the conformational space of IDPs. To the best of our knowledge, this is the first MetaD-based study of an IDP that offers such a quantitative evaluation of the atomic-level reliability of the conformational sampling via a very simple and intuitive metrics.

Two different schemes of population weighting were adopted. One was based on the reconstruction of the statistical ensembles following the Boltzmann equilibrium. The other was based on geometrical clustering and discarding scarcely populated ensembles. The former is derived from the FES binning and the latter from the structural similarity between N_{TAIL} conformers.

Sampling	Cα	Cβ	C'	HN	N
Clustering	0.887	0.451	0.752	0.213	1.863
Boltzmann	0.949	0.472	0.902	0.185	1.807
Kinetic reweighting	0.944	0.442	0.947	0.166	1.686
Sample and select	0.426	0.484	0.328	0.172	1.578

Table 1. Average Root Mean Square Deviations (RMSDs, in ppm) between Calculated (Using SPARTA+) and Experimental Chemical Shifts. Four different methods were compared, as indicated in *NMR data prediction* of the *Methods* section.

In both cases, there was very good reproduction of the experimental data, with deviations generally being very close to each other. This is a reflection of the small free-energy differences among the most relevant conformations (Figure 10). It suggests that the energetic convergence of the simulation (Figure 9) corresponds to an excellent sampling in terms of molecular geometry too. Interestingly, for the most structurally informative nuclei Cα, Cβ and C', the geometrical clustering provided a better agreement (in terms of RMSD,

see Table 1) than the FES binning. This can be explained in terms of suboptimal force field performance. Moreover, the great FES roughness, typical for IDPs,[105,106] can partially influence the accuracy of any binning/clustering procedure used for weighting the populations, hence randomly biasing the final result to a small extent. Furthermore, a sample-and- select procedure applied to the whole trajectory indicates that it contains the core of the chemical information encoded by NMR spectra. This indicates that MetaD can be efficiently used as a source of structural information in an SAS procedure stemming from the protein FES. Notably, the ability of the SAS based procedure to reproduce experimental data supports the use of MetaD to efficiently explore at least the full experimentally accessible conformational space of IDPs.

The absolute chemical shifts calculated according to the four different methods mentioned above are compare with the experimental counterparts in Figure 11. As expected, good correlations for Cα and Cβ atoms could be obtained, while a slightly noisy pattern was found for the carbonyls.

However, the most relevant probes for capturing details about the structure and dynamics along the sequence are the secondary structure chemical shifts (*ss* in the present text, i.e. the difference between the computed or measured δ and its value in the random coil).

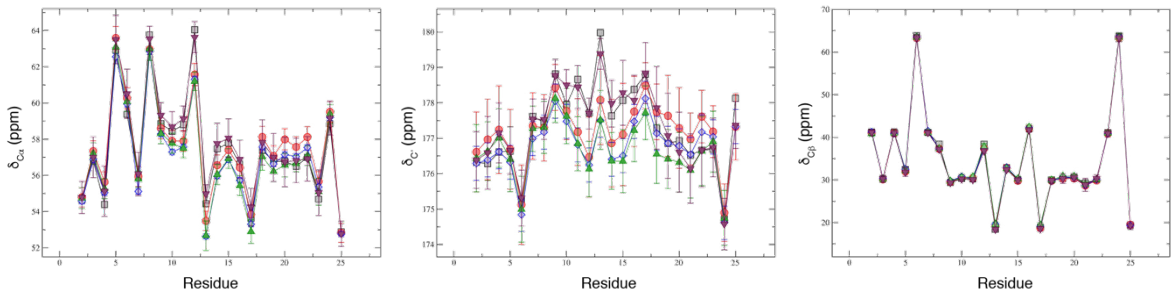

Figure 11. Absolute Chemical shifts for the Cα (left panel), C' (central panel) and Cβ (right panel) atoms of N_{TAIL}. Results from the calculations on the clustered trajectory (red, circles), the Boltzmann weighted trajectory (blue, diamonds), the kinetic ensembles (green, triangles) and the conformational selection procedure (maroon, triangles) are compared to the experimental values (black, squares).

In our calculations (Figure 12), $ss_{C\alpha}$ and $ss_{C'}$ exhibit positive values, indicating the formation of alpha helical structures throughout the chain, in agreement with the average slightly negative value of $ss_{C\beta}$. Notably, all the above-mentioned secondary shifts exhibit a significant deviation from the experimental data around residue Alanine 484, corresponding to the C terminus of helix H1 (479- 484). Several almost fully helical

structures displaying a tilt centred on residues 484-486 are present in the minimum energy region (see sample structures reported in Figure 10). This is compatible with the discrepancy between the simulated and experimental data. This tilt can be explained by the presence of a small three-residue hydrophobic cluster (I483, A484, and M485), with side chains that tend to stay closer in space in order to exclude the solvent, consequently inducing a higher mobility in the two branches of the chain. Most likely, more demanding simulative setup are required to reduce these discrepancies.[107]

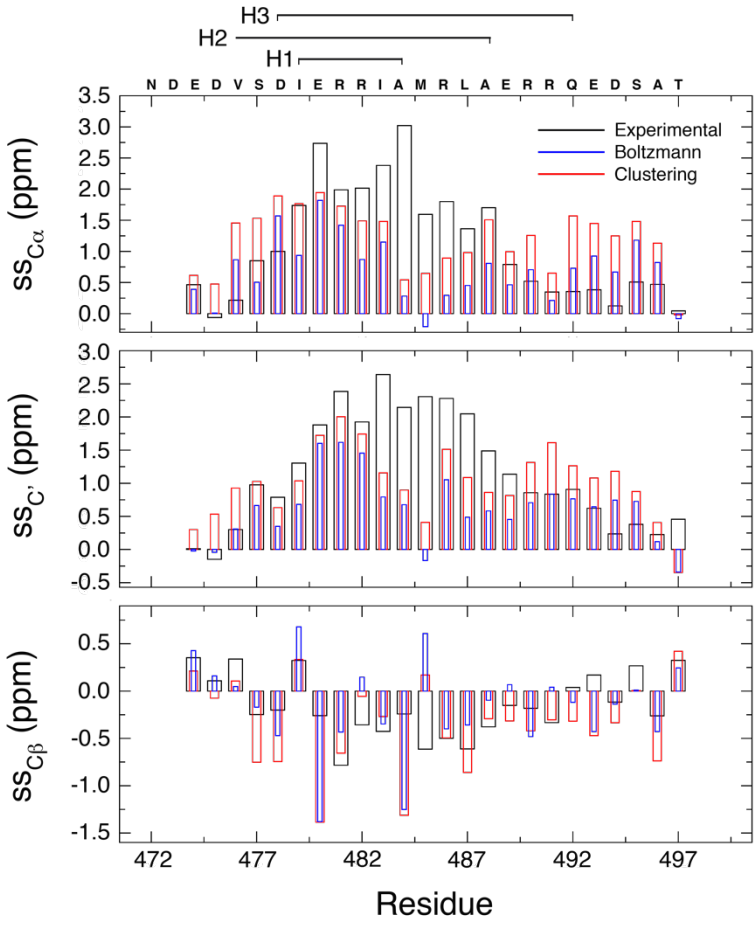

Figure 12. Secondary chemical shifts (difference with respect to the corresponding random coil values) for the Cα, Cβ, and C′ atoms. Boltzmann weighted (blue) and cluster weighted (red) values, calculated on the structures obtained from the MetaD simulations, are compared to the the corresponding experimental data (black). For clarity, the NTAIL sequence is also shown above the top panel, along with the α-helical regions in states H1, H2 and H3.

While previous experiments succeeded in providing structural information, little is known about the dynamics underlying the interconversion between accessible states, which is still a major challenge for biophysics. We used the statistics collected in our biased

simulations to construct a discrete- states kinetic model of the N_{TAIL} conformational ensemble in solution, following the approach in Ref.[26] The essential requirements for building the kinetic model are a FES and a matrix of diffusion coefficients in a CV-space that is not necessarily the same as that used during simulations. We obtained the former applying a reweighting scheme as explained in the *Kinetic model setup* subsection of the *Methods* section, obtaining the FES shown in Figure 13.

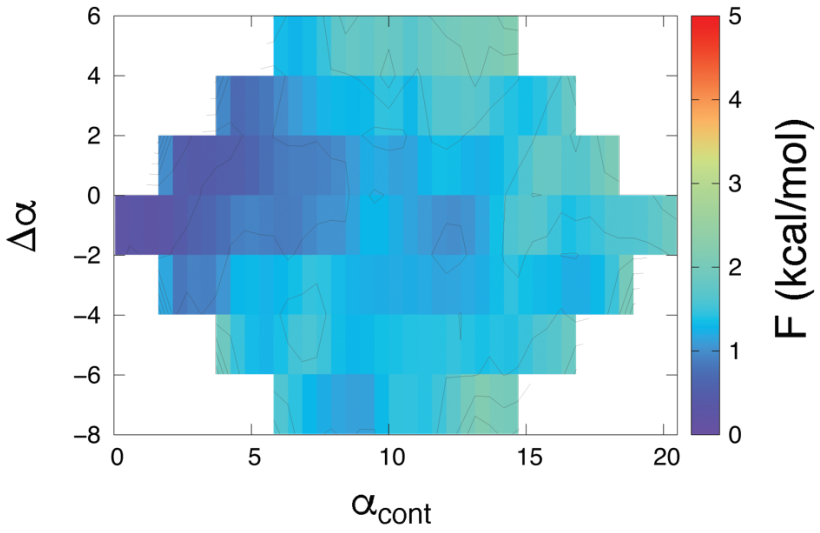

Figure 13. We employed all of the structures obtained from the MetaD simulation to reweight the FES in the two-dimensional space described by the α_{cont} and $\Delta\alpha$ CVs. To build the kinetic model, a binning procedure applying a 40 x 8 grid along the two CVs (α_{cont} and $\Delta\alpha$, respectively) was carried out.

To determine the latter, we maximized the likelihood that a given set of plain MD trajectories could be reproduced by the kinetic model, whose transition rate matrix is a function of the diffusion coefficient itself. The model allowed us to identify the kinetically representative conformations of N_{TAIL} and the corresponding rates (k_{ex}) between each pair of states. MFPTs are defined as the reciprocal of k_{ex}. Our kinetic model identified a total of six distinct states (states A to F in Figure 14).

Notably, all the relevant conformations previously identified in experiments[78] were also identified as major states in dynamic equilibrium in our model. In detail, the RC, H1, H2, and H3 conformations corresponded to states A, B, C, and D, respectively, while states E and F represented two additional conformations with a significantly high helical propensity. The numerical solution of the simplified system, as indicated in the *Calculation of equilibrium populations* subsection of the *Methods* section of this chapter, led to the following populations of the state in the stationary regime: PA = 30.5%, PB =

41.4%, PC =19.9%, PD =1.7%, PE =2.4%, PF =4.1%. A fairly good agreement could be observed between the obtained values for states A, B, C and D and the corresponding values for conformations RC, H1, H2 and H3 from the proposed conformational equilibrium in solution for N_{TAIL}.

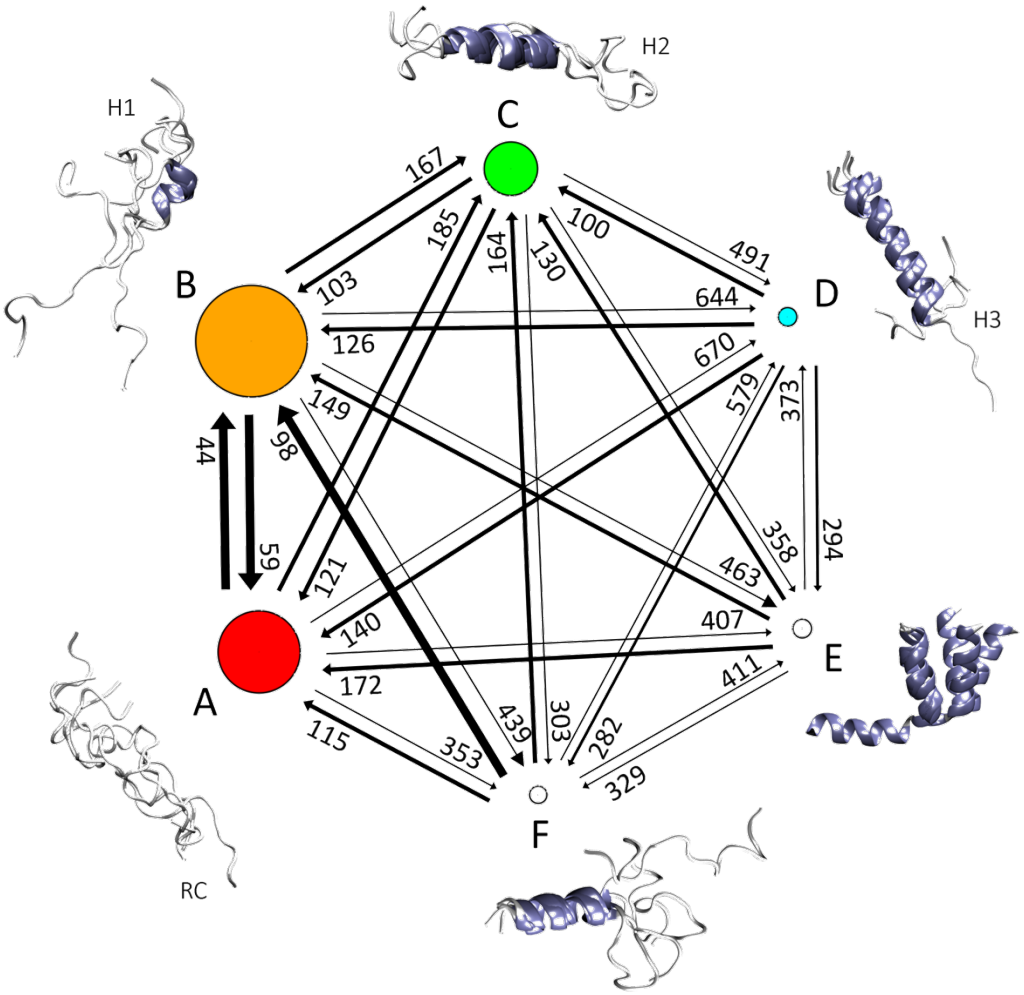

Figure 14. Interconversion between the kinetically representative conformations of N_{TAIL}. The areas of the discs are proportional to the equilibrium populations of the states, while the widths of the arrows are proportional to the corresponding MFPT. MFPTs (in ns) are indicated on the arrow ends, close to the destination state.

Remarkably, we computed kinetic rates that fall above, and are thus in agreement with, the regime of fast exchange (k_{ex} > 105 s^{-1}) indicated by previous works for the interconversion among equilibrium conformations for N_{TAIL} in solution.[13] A closer look at the kinetic model reported in Figure 14 allows two sub-ensembles of conformations to be identified, reflecting different interconversion regimes and thus distinct thermodynamic stabilities. The first group, characterized by the fastest transitions, with MFPT values

below 200 ns, is represented by states A, B, and C, corresponding to the mostly unfolded RC, H1, and H2 conformations, respectively. Conversely, interconversions between the remaining states (D, E, and F), possessing a higher alpha helical content, are significantly slower, requiring at least 280 ns (D to F transition) to take place. Going from the unfolded to the folded sub-ensemble through direct transitions is also much slower than the reverse. This is reflected by the consistently higher thermodynamic stability of states A, B, and C compared to states D, E, and F. Indeed, the slowest direct unfolding transition (E to A) is much faster than the fastest direct folding transition (C to F), with MFPTs of 172 and 303 ns, respectively. The binding of N_{TAIL} to its molecular target PX, the C-terminal domain of the viral phosphoprotein, is accomplished through electrostatically driven interactions.[108] The binding site is a negatively charged cleft on the surface of PX, with which N_{TAIL} interacts, adopting conformation H3. In this state, three arginine residues (R482, R486, R490, see Figure 6), located at the centre of the helical motif, are positioned on the same side of the helix, forming a positively charged counterpart of the PX binding site. It has been proposed that achieving the bound state involves a two-step process.[13] An encounter complex is first formed between PX and N_{TAIL} in the H2 conformation. This subsequently evolves into the native bound state by establishing specific electrostatic interactions. This framework suggests a coupled folding and binding process, arising from a sequential cooperation of the conformational selection and induced fit recognition mechanisms, as already observed in previous works on IDPs.[109]

In our model and in agreement with previous studies,[78] despites being favoured upon binding to PX,[13] state D (corresponding to conformation H3) is also shown to be accessible in the absence of the molecular target. In other words, even though it is intrinsically disordered, the N_{TAIL} sequence encodes for the structural determinants required to adopt the bound conformation. Moreover, among the unfolded sub-ensembles, state C (corresponding to conformation H2) can be easily accessed through both the A and B states, which are in rapid dynamic equilibrium. Remarkably, reaching C from state A is only slightly more favourable than from state B. This is interesting because it suggests there is no preferential route leading to the partially folded state C adopted in the encounter complex. It is accessible from the completely unstructured A without necessarily passing through state B.

3.1.4 Conclusions

In the present study, we used atomistic simulations to sample conformational states visited at equilibrium by NTAIL, an IDP test case,[77] free in solution. To the best of our knowledge, this is the first time that computational methods have been used to provide both thermodynamic and kinetic atomic-level details of an IDP with a quantitative comparison between experimental and spectroscopic data. In particular, we determined the free-energy landscape of NTAIL and calculated the kinetic rates for interconversion between the main free-energy attractors. Our study demonstrates how, despite the current limitations, a fairly good agreement with experimental data can be achieved, provided that there is an exhaustive sampling of the conformational space. This is encouraging, as it points the way to studies aimed at understanding the interaction with molecular targets, and eventually developing effective strategies to drug IDPs.[110] In particular, our findings suggest that MetaD is a particularly suitable methodology for disordered systems that present transient structures, also offering the possibility to explore their FES in a form that has a strong correlation with experimental data at atomic level.

3.2 TEST CASE 2: β2-AR

3.2.1 Introduction

A major goal in drug discovery is the identification of small molecule ligands that are able to bind and modulate the activity of biomolecules involved in a certain, or sometimes multiple, pathological conditions. Binding of small molecule ligands is usually represented as a two-state, all-or-none process in which the ligand is either free in the bulk or placed in a binding site on the molecular target. Ligand binding affinity and binding rates are quantitative parameters that are crucial during the drug discovery and development process. These parameters depend on the free energy profile of binding. Affinity is typically expressed in term of K_d and depends on the free energy difference between bound and unbound states, while kinetic rates, namely k_{on} and k_{off}, depend on the free energy difference between these states and the transition states, which are pretty elusive and difficult to observe (a schematic representation is given in Figure 15).

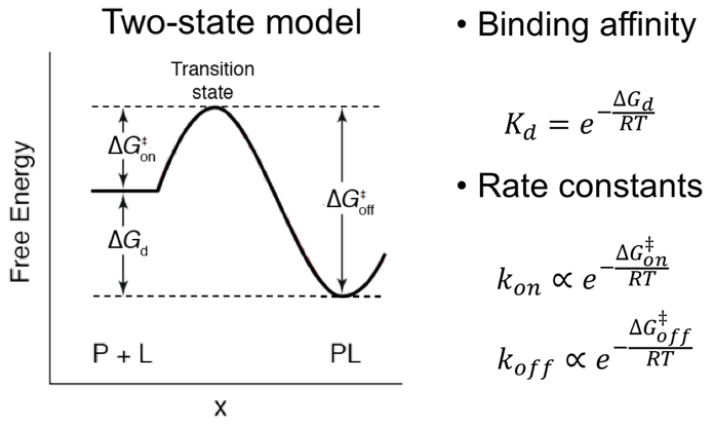

Figure 15. Schematic representation of a two-state model for the protein (P)-ligand (L) binding process. Assessing affinity (in terms of K_d) implies determining the difference in free energy between the bound and unbound states (ΔG_d), which in turn is independent from the route followed from the former (P + L) to the latter (PL) state. Conversely, calculating the on- and off- rates (k_{on} and k_{off}, respectively) requires a detailed knowledge about the intermediate states visited along the reaction coordinate (x).

In drug discovery and optimization phases, one wishes to achieve a rational optimization of these interesting quantities in order to improve drug binding properties.[4,6] Therefore, an accurate prediction of the free energy profile associated to ligand binding is of paramount relevance. However, this is still an ambitious goal for modern drug design. Additionally, a more complex picture compared to the two-state mechanism can be also envisioned, where a series of intermediate, lower energy states are separated by higher energy transition states. Complex reaction pathways are difficult to characterize at such level of detail through experiments. Therefore, computer simulations, and MD[24] in particular, can be extremely helpful in this regard as an all-atom description of complex processes, including protein-ligand binding, can be provided. In the limit of an ergodic simulation, useful insights about transition and metastable states can be extracted and the affinity and rates would be determined with high accuracy.

An increasing number of studies is demonstrating how simulating spontaneous binding/unbinding of small molecules to biomolecular targets via unbiased, plain MD simulations is becoming more and more viable,[111–114] also for systems of considerable size such as GPCRs.[111] However, the computational resources required are still impressively high, and certainly not accessible to most of the current research groups. Furthermore, in this scenario, collective proper statistics of such rare events is prohibitive. Enhanced sampling methods have been developed to overcome these limitations.[16] MetaD[25] is one of such techniques, and allows an efficient exploration of a system's phase space guiding the

sampling along reaction coordinates representing slow degrees of freedom, the so-called CVs. However, identifying appropriate CVs is not trivial and requires the major efforts when using MetaD.[31] Moreover, this aspect exacerbates when it comes to protein-ligand binding, where many, and potentially system-dependent, degrees of freedom, are involved. Thus, the path CVs[18] have been specifically devised to manage complex reaction pathways. The idea is to guide MetaD sampling along a putative pathway representing the progression along the reaction, while exploring adjacent regions of the phase space. While multiple pathways can be considered, the one in which the minimum free energy lies can be recognized. However, as one might expect, providing the required "guess path" is far from trivial, as this implies the availability of some sort of information about the process we are aiming at studying.

Herein, we present our strategy to tackle this problem. First of all, we construct a MSM[17] from available, large-scale plain MD simulations. The model allows identifying relevant intermediate states visited by the ligand along its way to the binding site. Subsequently, we exploit such states as a template to construct a putative binding pathway, as required by the implementation of the path CVs. Taking advantage of such "guess path", our aim is to reconstruct the free energy surface along the binding process through path CVs-based MetaD. This would allow us exploring the possible routes accessible to the ligand, determining the minimum free energy pathway, and identifying the intermediate metastable and transition states visited.

We applied our procedure to binding of the ligand Alprenolol to the β2-adrenergic receptor (β2-AR) (Figure 16). In a recent study, the D. E. Shaw research group reported 10 binding events for the ligand Dihydroalprenolol and 2 for the ligand Alprenolol.[111] While in three out of the ten runs from the first group the ionic concentration of Na^+ and Cl^- was specified in order to neutralize the system, in all of the other runs the necessary amount of Cl^- ions was added only. Accumulating the 12 runs provided a total simulation time of about 63 μs. Interestingly, 11 strikingly similar pathways were observed (see Table 2). Moreover, while in all of the simulations the ligand was able to reach the orthosteric binding site, in 6 of them only the binding pose reported in the crystal of the protein-ligand complex (PDB ID: 3NYA, resolution 3.16 Å) was reproduced (RMSD < 0.8 Å). Given the very small difference in chemical substituents (Figure 16, right-hand side), both ligands Dihydroalprenolol and Alprenolol showed the same interactions with the molecular target in the reported simulations as well as and in experimental studies. Therefore, the runs were

grouped and analyzed together in the original paper and the ligand was simply referred to as Alprenolol.

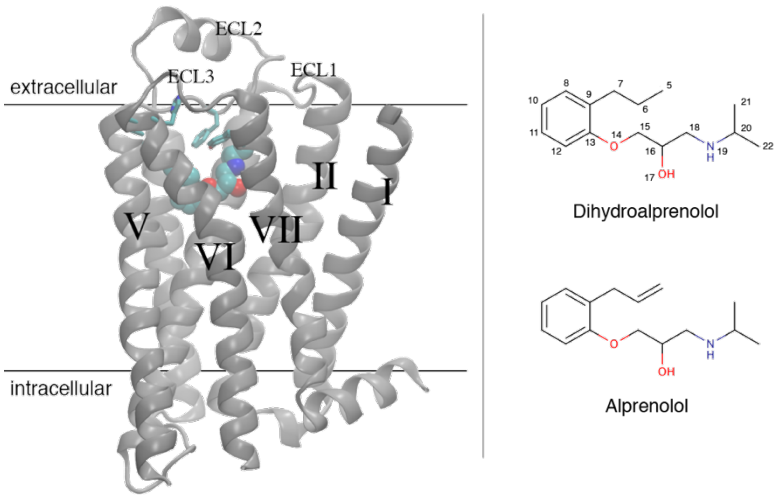

Figure 16. The β2-adrenergic receptor (β2-AR, left-hand side) and the two analogue inhibitors considered for the present work (right-hand side). The transmembrane helices towards the reader (I, II, V-VII) are labelled for clarity, as well as the extracellular loops (ECL1-3) between the helices; a schematic representation of the positioning inside the cellular membrane is also given. The location of the orthosteric site is highlighted by the presence of the ligand, shown in the van der Waals drawing method. 2D structures of Dihydroalprenolol and Alprenolol are shown on the right; according to the plain MD simulations and previously reported experimental data, no relevant difference in the behaviour of the two small molecules was observed. Numbering is shown on heavy atoms to help the reader when specific positions of the molecule are mentioned in the text.

Run	Ligand[a]	Condition[b]	Duration (μs)	Binding pose[c]
1	DA	N	11	-
2	DA	N	3	√
3	DA	N	5	√
4	DA	N	4	√
5	DA	N	10	-
6	DA	N	5	√
7	DA	N	9	-
8	DA	I	5	-
9	DA	I	3	√
10	DA	I	5	-
11	A	N	3	√
12	A	N	3	-

Table 2. Summary of the available trajectory set, as provided by the D. E. Shaw Research group. [a] DA stands for Dihydroalprenolol, while A for alprenolol. [b] N designates those systems that have been neutralized by adding only the appropriate number of ions of the opposite sign. I indicates instead that the ionic strength was set, thus both positively and negatively charged ions were present in the bulk. [c] Where a √ sign is present, the ligand was able to reach the binding pose reported in the crystallographic structure (PDB ID: 3NYA) with an RMSD < 0.8 Å.

3.2.2 Methods

We used all of the available simulations reported in Table 2 to build a MSM.[111] As previously mentioned, in only 6 out of 12 runs the authors were able to reproduce the binding pose observed in crystal structure of the complex (PDB ID: 3NYA) with an RMSD < 0.8 Å. We refer to these simulations as "productive". However, in all of the 12 runs the ligand reached the orthosteric site on the β2-AR. In other words, while in the remaining cases Alprenolol was able to access the binding site, it remained stuck in alternative poses without reproducing the crystallographic one. These are referred to as "non-productive" runs. Therefore, we aggregated both productive and non-productive runs to construct the model, as information about access to the orthosteric site is guaranteed in both cases. In order to reach the site, which lies buried deep inside the bundle of seven transmembrane helices, the ligand needed to squeeze through a narrow passage between

ECL2 and and helices 5-7. To approach the passage, the so-called extracellular vestibule is first occupied, enclosed by ECL2, ECL3 and helices 5-7.

All of the simulations were started in the unbound state, with 10 replicas of the ligand placed in the bulk. Trajectory frames were saved every 180 ps, and included a significant amount of time during which ligands wandered in the solvent without accomplishing durable contacts with the solvent-exposed surface of the protein. Therefore, in order to reduce the amount of data to process, and considering that our goal was to achieve a thorough understanding about the behavior of the ligand from the surface to the orthosteric site, we discarded these initial, non-relevant parts. Specifically, for each trajectory, we defined contacts between the Alprenolol and residues from the β2-AR when the distance between heavy atoms was below 5 Å in at least one frame. We then monitored the distance relative to such contacts along the trajectory, and discarded those frames in which the minimum among the distances recorded was above a 30 Å value. Applying this procedure, we gathered a total number of frames equal to 296659.

In the upcoming sections, we first go through the steps followed to build the MSM for Alprenolol binding to the orthosteric site of the β2-AR. Subsequently, we highlight the procedure applied to select relevant states along the binding pathway. In order to carry out these steps, the software PyEMMA, version 2.4, was used.[115] Finally, we describe how we constructed the guess path.

3.2.2.1 Choice of the variables

As already pointed out in the *Theory* chapter, the first step when constructing a MSM is selecting appropriate variables.[37] By using the term appropriate, we refer to variables that are able to describe the dynamics of interest, and to recognize different states of the system relatively to the process under study. In this specific case, the aim was to monitor the advancement of Alprenolol along its route from the surface of the protein towards the occupation of the orthosteric site.

A first obvious choice would be considering the RMSD of the ligand after aligning the system on the protein α-carbons. However, distinguishing such a plethora of different states taking advantage of a single value can be misleading. In particular, similar values of RMSD could gather together extremely different configurations of the protein-ligand system. Therefore, after considering such option and confirming the expected behavior, we

discarded this choice. Another obvious possibility is the distance between the center of mass of the ligand with respect to the one adopted in the crystallographic binding pose. However, similar issues as for the RMSD were encountered in this case. Specifically, different configurations of the ligand can be expected at same values of distance. We iterated over several other options, and tested the outcomes in the ability of clustering to group together similar states employing different variables and combination of variables. Inspired by a previous work in which the authors tackled similar issues,[116] we considered the minimum distance between protein and ligand heavy atoms. Another aspect, when choosing variables, is the desired trade-off between their number and the ability to be good variables. Since clustering is going to be applied, the wider the variable space, the highest the computational effort required. Moreover, in our experience, using increasing number of variables would not necessarily lead to an improvement in the definition of different states, and we related this effect to an increase in the overall noise. Therefore, we reduced the picture to considering α-carbons only for the protein, and heavy atoms for the ligand. Moreover, since the process under investigation involves directly only a fraction of the protein residues, we restricted our selection to those belonging to the bulk-exposed surface, to the extracellular vestibule, to the narrow passage leading to the binding pocket, and to those making up the orthosteric site. Figure 17 gives a pictorial representation of the 64 protein residues resulting from this selection. As for the ligand, not all of the heavy atoms were considered as well. We reduced the selection to the most significant heavy atoms needed to describe a relevant configuration of the ligand. Specifically, the two oxygen atoms belonging to the phenol ether and the hydroxyl groups respectively, and the nitrogen present in the secondary amine group, positively charged at the physiological pH. These chemical groups allow the ligand to interact with protein residues by means of hydrogen bonds and salt bridges. Additionally, two carbon atoms were selected in order to take into account the possibility of having a flip of the phenyl ring while the other chemical groups in the small molecule would maintain the same configuration. Following the numbering reported in Figure 16, the 5 atoms O14, O17, N19, C11 and C7 were chosen. Therefore, 320 minimum distances between these two groups of atoms, namely the α-carbons from the 64 protein residues selected and the 5 heavy atoms from the ligand, were obtained and used as variables for the next steps.

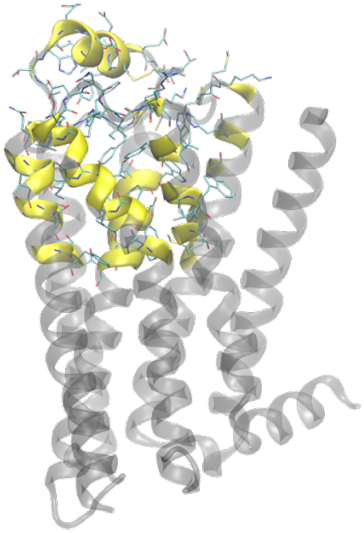

Figure 17. Protein residues (highlighted in yellow, all of the corresponding heavy atoms are shown in sticks) selected for the calculation of minimum heavy atom distances with respect to the ligand. From these residues, the α-carbons only were considered.

3.2.2.2 Clustering of the variables

After appropriate variables have been defined and their values calculated for each frame of the available trajectories, the subsequent phase involves clustering such data. We call this step state decomposition,[37] meaning that we are dividing the system's configurational space into a set of states (the clusters), referred to as microstates. Since the aim is to quantify the transitions between these microstates, in the ideal situation one would achieve a kinetic clustering. This would require determining the average transitions times between all couples of conformations visited by the system, and defining groups basing on such information. In such a scenario, no internal energy barriers would be encountered within a microstate and all of the members would interconvert more rapidly between them than with respect to any other microstate. However, there is currently no way to obtain these average transition times easily and in a straightforward manner. Therefore, the typical procedure is to devise a kinetically relevant geometric clustering. The idea is that similar configurations in terms of geometry are likely to rapidly interconvert. On the one hand, this would become even more likely as we divide the configurational space more and more finely. On the other hand, we need to ensure sufficient statistics for each microstate, or, in other words, a reasonable population for that cluster, in order to determine transition probabilities between the microstates accurately.

In order to cluster our data, we employed the K-means clustering algorithm.[43,44] We emphasize that several clustering algorithms have been devised over the years,[117] but no general algorithm exists that one can expect a priopri to be effective. Clustering is not an application independent problem, it is strictly dependent on the context, and each algorithm has its own advantages and disadvantages. As for K-means, it has gained great popularity in the MSM community and has been employed with satisfactory results in several previous works.[40–42] This clustering algorithm tends to locate more clusters in more densely sampled regions. This avoids creating too many groups with small counts, thus favoring a more accurate determination of the transition probabilities. However, one needs to be aware that risks are the over division of certain regions and under division of others. When using K-means, the number of clusters needs to be specified by the user. A rule of thumb is to use the square root of the number of initial data points. In this case we had about 300000 data points, thus resulting in a number of clusters of 500. As stressed in the original paper,[111] strikingly similar pathways were observed in the binding trajectories. Visual inspection confirmed such picture, and we observed a contained amount of states in which the ligand stations for a long time. Therefore, we considered reducing the number of cluster to as low as possible. We tested increasing number of clusters, specifically 50, 100, 150, 200, 300, and 500. Since we worked in a 320-dimensional space, we projected the corresponding clusters in a more intuitive space in order to assess the quality of the results. For each cluster obtained, we calculated the RMSD on the ligand with respect to the crystal bound state for all of the configurations contained in that cluster, and took the average; similarly, the average distance between the center of mass (COM) of the ligand in each configuration of the cluster and in the crystallographic complex was used as a second dimension. In Figure 18, the different numbers of clusters considered were projected in such space. As shown by the plots, increasing the number of clusters from 50 to 100 and 150 clearly leads to an improved distribution of the clusters. Conversely, a further increment to 200, 300 and 500 does not broaden significantly the distributions, while determines instead a finer subdivision of the already considered space. As already introduced, we aimed at maintaining the amount of clusters contained. Therefore, we chose 150 as the number to build our model.

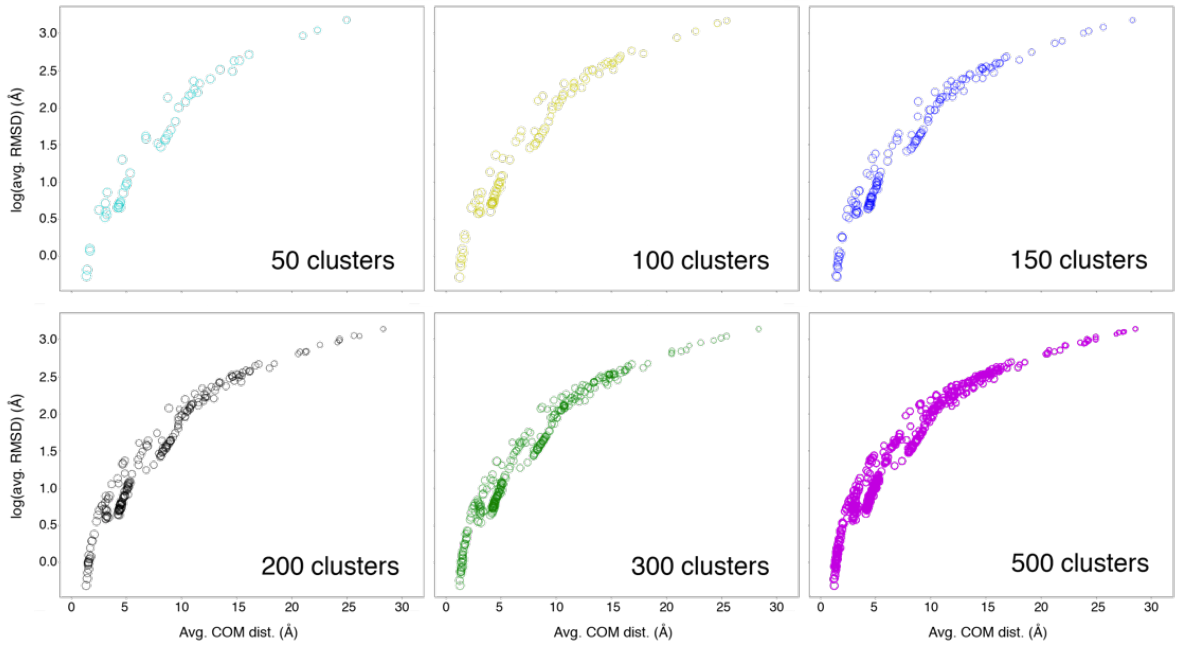

Figure 18. Projection of different number of clusters in a 2D space. Distributions of the clusters in the considered space do not change significantly increasing their number from 150 to 200, 300 and 500.

3.2.2.3 Lag time selection and Chapman-Kolmogorov test

Constructing a MSM can be roughly summarized in two major stages. First of all, a state decomposition is carried out to define the microstates in the system. This allows shifting from the conventional view of trajectories as a series of structures over time to a series of microstates over time, the so-called discrete trajectories. Secondly, jumps are performed over these discrete trajectories and the corresponding transitions in the microstate space recorded and stored in a count matrix, from which a transition probability matrix can be eventually derived. A critical element in this picture is the size of such jumps, as different count matrices would be obtained as a consequence. As already introduced in the *Theory* chapter, we refer to such element as the lag time of the model. Markov time is called the smallest lag time that gives the Markovian behavior.[37] Under the Markovian assumption, systems are memoryless, meaning that the probability of being in the current state only depends on the previous state, and not on all of the preceding ones. In this view, transition counts are statistically independent. Assessing the lag time dependence of the relaxation time scales implied in the system dynamics has been shown to be a useful approach to determine appropriate lag times to ensure memory loss. The outcoming plot is usually referred to as ITS plot (Implied Time Scales plot).[37,38]

We monitored the dependence of the time scales of the system on increasing values of lag time. The state decomposition phase was repeated three times and the corresponding ITS plots determined. As shown in Figure 19, very similar results were obtained. As can be observed, the relaxation time scales level off at a value of about 600 steps. As already mentioned, when performing the simulations that we subsequently exploited for this work, the authors saved trajectory frames every 0.180 ns. Since we did not stride the trajectories, 1 step corresponds to 0.180 ns. Therefore, our choice of 600 steps corresponded to a lag time of about 100 ns.

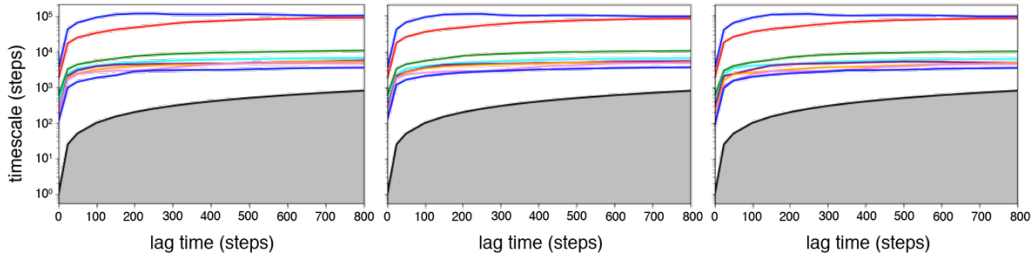

Figure 19. Relaxation time scales as a function of lag time (ITS plots). The results after performing three independent state decompositions are reported. In all of the three cases, lag time dependence of the time scales ceases after a lag time of about 600 steps.

Notably, in order to carry out the procedure just described above and to produce the plots in Figure 19, our first attempt was to aggregate all of the 12 productive and non-productive trajectories from the original paper.[111] However, this did not allow observing clear convergence of the relaxation time scales. As such a behavior would typically indicate presence of states not adequately sampled, we run through the simulations to identify possible sources. Thus, while in 11 out of the 12 simulations the binding pathways were extremely similar, we observed one single trajectory in which the ligand got access to the orthosteric site through a different route. Therefore, we discarded the trajectory and repeated the analysis. This allowed achieving a significant improvement in the ITS plots, obtaining the converged plots reported in Figure 19. While this clearly indicates that the procedure is extremely sensible and responsive to the input data, we are aware that we neglected part of the information. Nevertheless, we do not exclude that other binding pathways might be envisaged. Therefore, we decided to focus our view on the most frequent route observed in the available simulations without producing additional plain MD data.

One validation that is typically considered at this point is the Chapman-Kolmogorov test (CK test).[38] Through this procedure, we evaluate whether the MSM, obtained imposing a certain lag time on a discretized phase space, is consistent with the data used to parameterize it. In practice, what one does is assessing the probability of remaining within an initial microstate after multiples of the lag time used to construct the model. This procedure is repeated for MSMs built at increasing values of lag time, and for the original MD trajectories. The aim is to determine whether consistency is met between the MSM built at the desired lag time and the MD data within statistical error. Herein, we carried out the analysis constructing three MSMs, specifically with lag time of 200, 400 and 600 steps. As can be seen in Figure 20, an increasing consensus with the plain MD data was found. In particular, the MSM built with the chosen lag time of 600 steps lies within the uncertainties of the transition probabilities estimated from the MD trajectories.

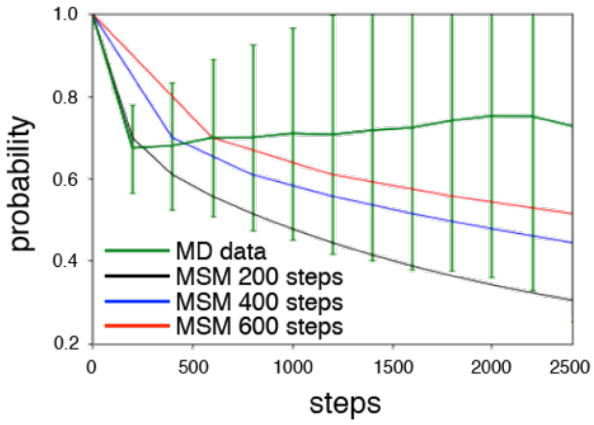

Figure 20. Chapman-Kolmogorov test. The consistency between MSMs constructed at increasing values of lag time (200, 400 and 600 steps) and the original MD data (green line, with bars indicating the standard error) is evaluated by means of the test.

3.2.2.4 Selection of the relevant states

We exploited transition path theory (TPT)[118,119] to identify relevant states along the binding pathway of Alprenolol from the protein surface to the orthosteric site. In order to apply TPT, an initial and a final state, typically denoted by the *A* and *B* letters, need to be indicated. At this stage, we employed the ligand RMSD with respect to the crystallographic binding pose. Thus, we chose as initial state the one with the highest RMSD, and the one with the lowest as the destination state.

We carried out pathway decomposition in order to obtain all possible pathways leading from state *A* to state *B*. For each pathway, the associated flux was also determined. Subsequently, we devised a simple prioritization procedure for the MSM states visited in all of the pathways obtained. Specifically, each state was assigned a score basing on the number of times it was observed in the ensemble of possible pathways, weighting for the fluxes associated to these pathways:

$$fs = \sum_{i=1}^{N} f_{AB,i} / f_{AB,tot} \tag{51}$$

where the summation runs over all of the possible pathways *i* in which the considered state is present, $f_{AB,tot}$ is the total flux related to the transition *A* to *B*, and $f_{AB,i}$ is the flux associated to pathway *i*. Through this analysis, the top ranking states were identified. It is worth noticing that centroids are defined by the K-means clustering algorithm as averages over the variables values possessed by all of the states gathered in each cluster. While it has the clear advantage of generating centroids that are representative of the clusters, this does not necessarily correspond to physically meaningful states. Therefore, the mdtraj software[120] was employed to extract representative configurations from the top ranking microstates. Specifically, those structures possessing the lowest RMSD with respect to any other one contained in the same cluster were selected as centroids.

3.2.2.5 Construction of the guess path

In order to use the path CVs, a frameset representing a putative pathway needs to be provided.[18] The series of frames catches the system at intermediate states along the process of interest, namely ligand binding in the present case. An essential requirement is equal spacing in terms of RMSD between these subsequent snapshots. Accordingly, including more frames has the effect of reducing such distance, thus increasing the resolution of the output free energy.

Obviously, few states would not be sufficient to reconstruct an accurate FES of the complex process we were facing. Considering the states that we extracted basing on information from the MSM, the RMSD distances between subsequent frames were respectively 5.8 Å, 7.7 Å, 7.2 Å, 1.2 Å and 2.5 Å. First of all, significant distances were present amongst some of the states, thus inevitably affecting the resolution of the FES. Secondly, they did not guarantee the equal spacing required. Therefore, we devised a

multi-step procedure in which we first enrich the amount of configurations and subsequently select equally spaced frames.

In order to achieve the desired enrichment, we performed a series of steered MD[28,29] simulations starting in each one of the identified states. As parameters for the protein and the ligand could not be transferred from the original topology, the entire system was re-parameterized. Both protein and lipids were modeled according to the Amber ff14,[121] in which Lipid 14[122] is comprised to describe the latter. Ligand parameters were taken from the general Amber force field (GAFF)[123] and charges were assigned following the typical RESP procedure.[124] In each of the steered MD runs between two of the subsequent states identified through the MSM, the RSMD was used as CV and targeted to 0. Moreover, one additional run was carried out from the first point. In this case, we acted increasing the coordination of the ligand by water molecules in order to obtain additional points with fully solvated states. We then reverted this first short trajectory and merged it with the other steered MD runs, resulting in a 30 ns-long binding trajectory. For the resulting 6 runs, 5 ns-long each, force constants ranging from 20 to 50 kcal/(mol Å^2) were applied. The work required was monitored and each run was repeated multiple times to ensure reproducibility of the outcomes. As expected, since ligand configurational space from the surface to the orthosteric site was relatively limited, all the RMSD-based runs were extremely similar. Sample work profiles from the runs are reported in Figure 21B. Conversely, as shown in Figure 21A, increasing the solvation of the ligand while detaching from the protein opened the way to a variety of states, less relevant for the purposes of our study. Therefore, amongst the runs characterized by lower work, we selected in this case the one in which it was minimum. All simulations were carried out by means of the GROMACS MD engine, version 4.6.7,[93] patched with the PLUMED software, version 2.1.1,[94] in order to perform steered MD.

Once the 30 ns binding trajectory was obtained, we proceeded to selection of equally spaced frames in terms of RMSD. Our aim was a spacing of about 1 Å, thus allowing to distinguish relatively close states along the binding pathway and to achieve a reasonable resolution of the outcoming FES. This was a non-trivial step, as required iterating until the initial amount of 13000 frames was reduced to a more contained number, and the desired spacing was achieved at the same time. At this stage, we took advantage of a script wrote by the path CVs developer, specifically devised for this purpose. Once an input trajectory with *N* frames is provided and the user specifies the number *M* of output states, the script iterates until the best selection among the available frames is found.

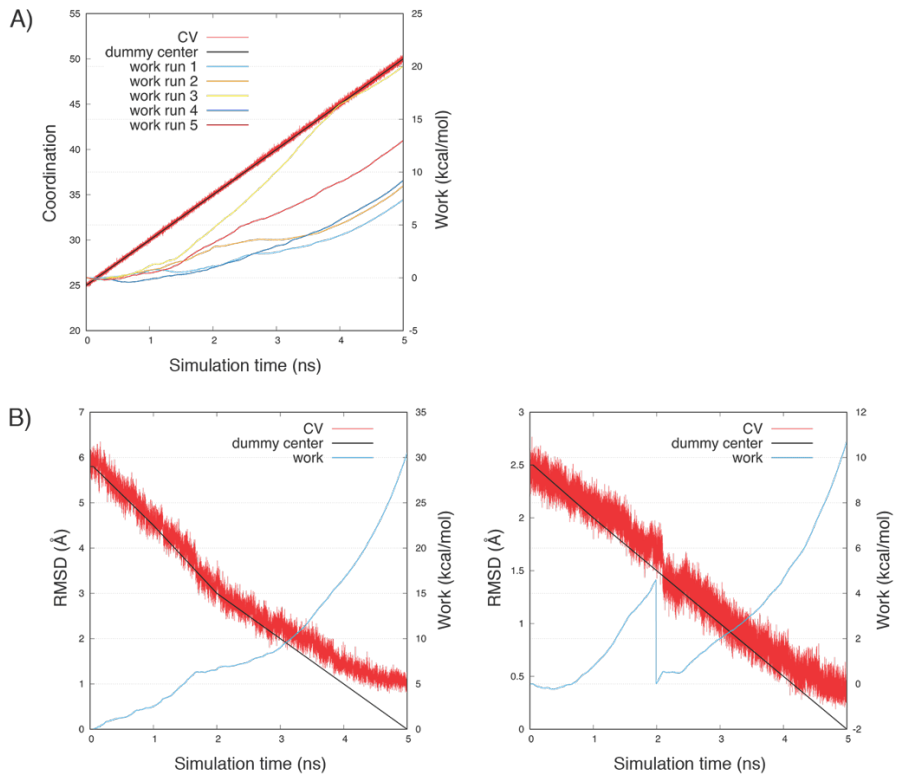

Figure 21. Work profiles from the performed steered MD simulations. A) Ligand coordination by water molecules was increased during the first step, in order to enrich with completely solvated states of Alprenolol. Different profiles (work run 1-5) were obtained, thus the one requiring the least work was retained for subsequent analysis. B) First (left panel) and last (right panel) steered MD runs using as CV the RMSD with respect to the subsequent state.

Figure 22 compares the RMSD between subsequent frames obtained from the first and the last iteration steps. As shown in the x axis of the figure, this allowed us selecting 80 states separated by 1.25 Å.

It is worth noticing that a requirement for the script is that the states in the input trajectory need to be already sequential. However, this is not obvious, as even in a steered MD run the system is able to fluctuate around a specific region of phase space and revisit it. This was particularly likely in our case, in which we considered the least aggressive force constants as possible to guide the ligand. Therefore, we devised an in-house script to solve this issue and remove from the steered MD simulation the mentioned loops. After applying this procedure, we gathered 110 frames that uniquely represented the advancement of the ligand in the 30 ns-long steered MD trajectory. However, such amount was not sufficient to guarantee that equally spaced, subsequent frames were present and could be identified by the method described above.

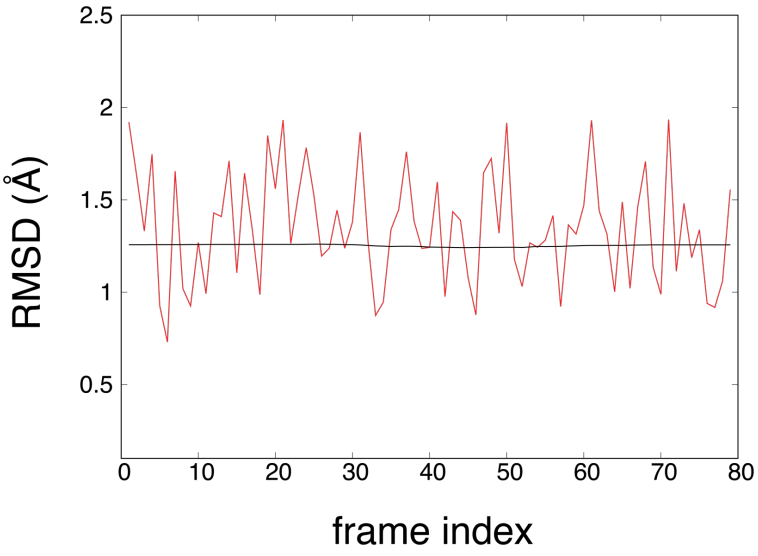

Figure 22. RMSD between subsequent structures calculated over the 80 frames comprised in the frameset. At the beginning of the procedure, 80 among all of the available frames were selected. This initial arbitrary choice returned significantly different values of RMSD between subsequent frames (red line). After a certain number of iterations, frames for which the spacing was as close as possible were identified (black line).

Therefore, we took advantage of a well-established interpolation method, the Catmull-Rom splines,[125] to use the 110 frames as control point and add intermediate frames in between. It is not our purpose to go into details about this interpolation technique herein, interested readers are exhorted to find more exhaustive material in the literature. The outcome of this pipeline was a set comprising 1091 frames, to which we applied the procedure described above in order to extract the guess path made up of the 1.25 Å-spaced 80 states.

3.2.3 Results and discussion

We aggregated available plain MD simulations for Alprenolol binding to the β2-AR in order to build a MSM of the process. To this end, as already introduced in the *Methods* sections, we selected 11 out of the 12 binding simulations as extremely similar pathways were observed. In order to construct a count matrix of the transitions in the discrete trajectory space, we selected a lag time of 600 steps from the ITS plots. Since 1 trajectory step is equal to 0.180 ns, this corresponded to a lag time of about 100 ns. By storing the transitions in the count matrix and subsequently determining the corresponding transition probability matrix, we built our MSM for the protein-ligand binding process. Starting from

the 150 microstates defined through clustering in the state decomposition phase, 93 % was the fraction of retained states in the MSM, corresponding to 140 states. It is worth pointing out that, despites the aggregated MD trajectories reached a simulation time of about 60 μs, the corresponding sampling for the binding process is extremely limited, as 12 events do not guarantee adequate statistics for such a complex process, for which multiple pathways might also be envisaged. Moreover, runs were stopped once the ligand reached the crystallographic pose, thus no unbinding neither re-binding were observed. Nevertheless, taking advantage of the 11 similar pathways, we were able to construct a non-reversible model for the considered binding process. The obtained MSM is shown in Figure 23, projected in a 2D space.

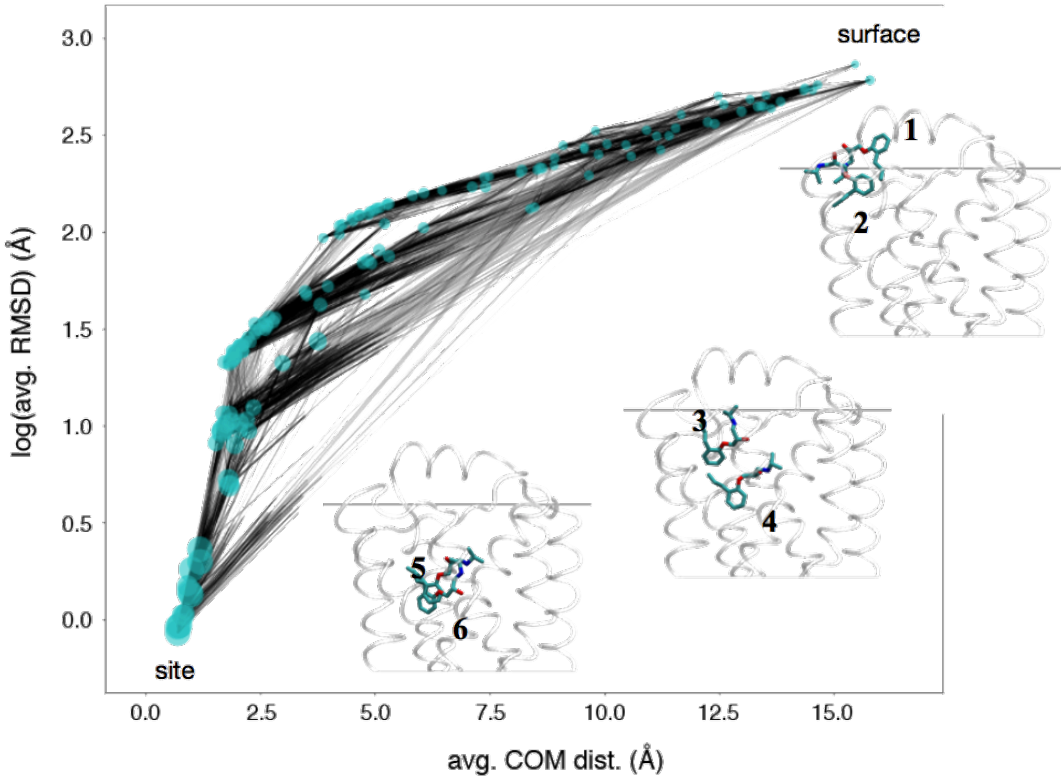

Figure 23. Markov State Model obtained for Alprenolol binding to the β2-AR. Circle sizes are proportional to the MSM microstate population, while black arrows indicate possible transitions between the states. The 6 structures in the insets represent the relevant states along the ligand route from the protein surface (top right-hand side of the model) to the binding site (bottom left-hand side) that have been selected by means of TPT, according to the procedure described in the *Methods* section.

Notably, each microstate was defined by 320 variables, representing minimum distances between heavy atoms from specific protein residues and from the ligand. Thus, in order to interpret the outcome more easily, we projected the model in a more intuitive

space. Specifically, we calculated the average RMSD of the ligand with respect to the crystallographic binding pose from all of the configurations contained in each MSM microstate, and, similarly, we took the average distance between the center of mass (COM) of the ligand in each configuration of the MSM states and in the crystallographic complex. Configurations in which the ligand was located on the protein surface were found on the top right-hand corner of the plot, while, on the bottom left-hand one, MSM states were located in the orthosteric site. As can be observed, at same distances along the binding pathway, different values of RMSD are possible. Clearly, this meant that the ligand was able to station in same regions while adopting significantly different configurations.

Following the strategy described in the *Methods* section, MSM states that were visited more frequently by the ligand on its way to the orthosteric site were identified. Shown in the insets of Figure 23 are the corresponding configurations, which were subsequently employed as a template to construct the guess path. Remarkably, all of the metastable states recognized in the original paper[111] were caught following our procedure. Specifically, such agreement refers in Figure 23 to poses 2 and 3, in which the ligand approached and occupied the extracellular vestibule, and pose 4, located at approximately the half way between the extracellular vestibule and the orthosteric site. Interestingly herein, besides pose 1, that represented an arbitrary state on the surface, almost completely surrounded by water molecules, we identified pose 5, in which the hydroxyl group was captured while reorienting from state 4 towards the final bound state 6. It is worth noticing that specific interactions stabilized each of the states identified. This was in support of their relevance in the binding process.

The 6 states were employed as starting points for the construction of a guess path. As already stressed, the latter was required in order to exploit the implementation of the path CVs. The resulting putative pathway, that is essentially a frameset describing a likely binding pathway according to the input data, comprised 80 equally spaced structures. As a result of the RMSD matrix optimization procedure, the distance between each subsequent structure was 1.25 Å in terms of RMSD.

In order to validate the quality of the reconstructed guess path, we projected all of the plain MD trajectories used to build the MSM in the space defined by the path CVs, namely the *s* and *z* variables. The result is shown in Figure 24, where different colors in a scale going from black to light gray identify the different input trajectories. As already introduced, through the *s* variable we monitor the progress of our system along the frameset. In other words, at a certain simulation step, the value returned by this variable is

going to identify the index of the closest among all of the available frames in the guess path. Therefore, since we had 80 structures in the guess path, the allowed values went from 1 to 80. The *z* variable, instead, is a measure of the distance from the frameset, calculated in terms of MSD and thus expressed in $Å^2$. For a more intuitive interpretation, we showed the square root of *z* in Figure 24. For easier reading, from now on in the text, we are going to discuss about *z* values in terms of Å, implicitly referring to the square root of the real values of *z*. In practice, for a certain state of our system, with a combination of the *s* and *z* variable we determine respectively which frame in the frameset is the most similar to the current state and quantify how actually similar is the system with respect to that frame. The purpose of this strategy is to also take into account states that are relatively far from the guess path, and, since the associated free energy is characterized by MetaD, to detect the minimum free energy route for going from frame 1 to 80. According to the projection showed in Figure 24, most of the plain MD data was located at low *z* values. This indicated that the constructed guess path represented very well the regions of the phase space visited by the original data. As expected, part of the data was placed farther away, reaching values of *z* as large as 7 Å and up to 12 Å. While, as already stressed, the pathways leading to the orthosteric site were extremely similar, some of the trajectories did not achieve the crystallographic binding pose and featured different states in which the ligand stationed for a significant amount of time. Indeed, the regions observed at higher values of z reflected such behavior.

Moreover, we also projected on the path CVs space the 6 states that were used as a template to construct the frameset. They are shown as larger white dots in Figure 24, and are numbered accordingly to Figure 23. As can be observed, all of the 6 states lied very close to the guess path, as they were characterized by low values of the *z* variable. This finding was very encouraging, as indicated that the guess path was well parameterized on the states used as a template.

In order to give a quick overview and interpretation, we subdivided the plot in two major areas. A left side that comprised the regions labeled as bulk and surface, and the right side that included those indicated as extracellular vestibule and site. Points belonging to the left side were more spread. Before achieving contacts with the extracellular vestibule, the ligand lacked of stable interactions with protein residues and was from completely to partially solvated by water molecules. As a result, the accessible phase space was extremely wide and no specific configurations appeared to be particularly favored. Conversely, as the extracellular vestibule was approached by the ligand, the configurations

were more concentrated in specific regions. Moreover, once the orthosteric site was reached, most of the trajectories lied very close and within a 4 Å distance from the guess path, highlighting high similarity in the visited phase space.

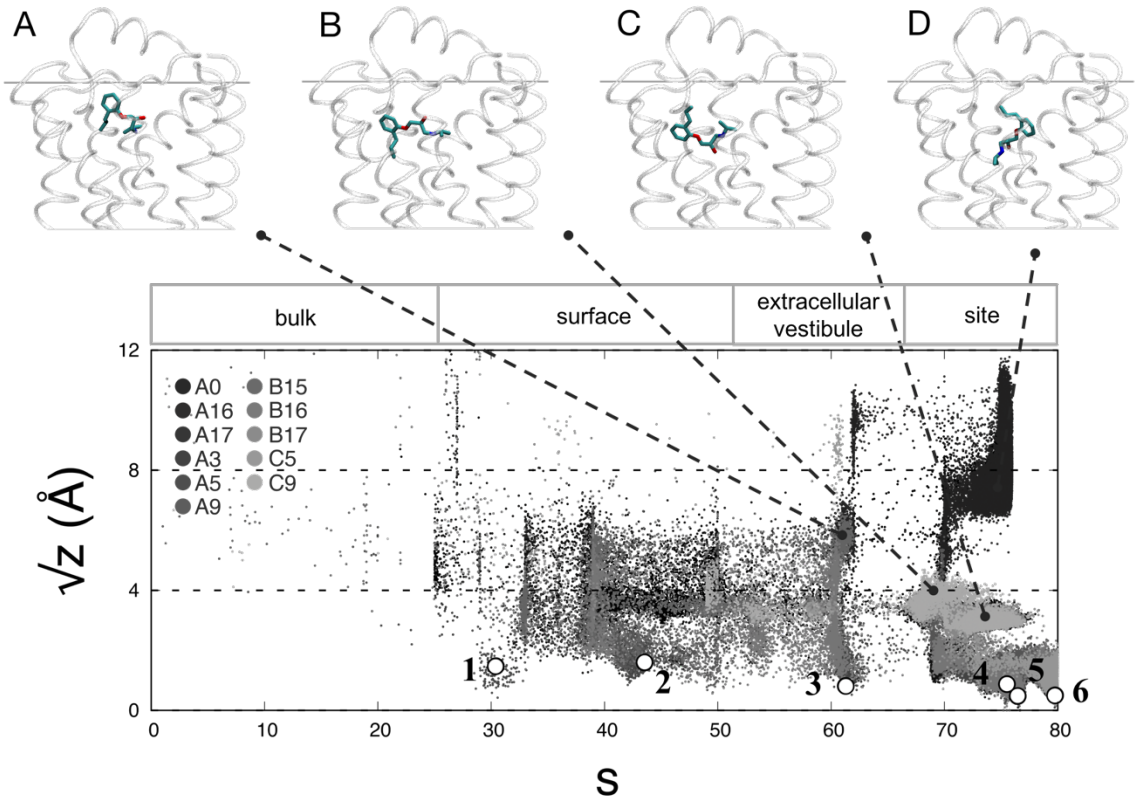

Figure 24. Projection of the original plain MD trajectories on the path CVs space. For an easier interpretation, the square root of *z*, in Å and equivalent to RMSD, is shown on the y axis instead of the real MSD values in Å^2. Right above the plot, a gross indication of the ligand position (bulk, surface, extracellular vestibule, site) in correspondence of *s* values is given. Each colour, from black to light grey, identifies one of the 11 different trajectories used to build the MSM. The larger white dots, labelled by numbers from 1 to 6, are the projection of the 6 states used as templates to construct the frameset. At the very top, structures labelled by letters A, B, C and D are shown for sample, more concentrated regions of the CVs space.

Shifting to a more practical perspective, the *z* variable is typically assigned an upper wall during MetaD simulations, meaning that sampling is confined within a certain value along this variable. In particular, in previous applications, it has been common practice to allow for a maximum value of 3 Å (corresponding to 9 Å^2 for the real values of the variable). For the purposes of this work, in order to include the largest amount of sampled regions in plain MD, a value of 4 Å would be more appropriate. While one might be tempted to include larger regions of the CVs space, allowing for sampling to larger *z*, this

choice is not free of drawbacks and inevitably exhibits pathologies. First of all, this obviously would translate into a significantly more demanding setup in terms of computational cost. Secondly, approaching higher values of *z*, we start loosing the resolution that we aim at achieving in the resulting free energy surface. Focusing our attention on the regions with *z* higher than 4 Å in the plot reported in Figure 24, we can observe that points tend to be compressed and to form narrow lines in the CV space. This is an unavoidable behavior, as the consequence of going significantly farther away from the reference path is that different conformations are going to be assigned to a same s value, thus creating the observed lines.

In order to assess the effect of limiting sampling at a maximum of 4 Å for *z*, we back-tracked the states belonging to the more concentrated regions observed in the plot. Sample corresponding configurations are shown explicitly at the top of Figure 24. In particular, state C, that we expected to be a reasonable state to be visited along the way from states 3 to 4 (highlighted by the white dots and labeled with the corresponding number, accordingly) lied within the 4 Å range that we aimed at considering for our production phase. State B, where the ligand is flipped completely with respect to the orientation assumed in the guess path, would be also considered with such setup. This was particularly exciting in the perspective of possibly including, in the outcoming free energy, pathways that differ significantly from the parameterized one. States A and D, also very different from the configurations included in the guess path, lied instead over the 4 Å cutoff. As a consequence, by employing such cutoff, these sates would reasonably be excluded from sampling in the subsequent MetaD production phase.

The above setup is going to be exploited to reconstruct the free energy surface of the protein-ligand binding process, namely Alprenolol binding to the β2-AR. To this end, path CVs-based MetaD is going to be performed. From the free energy landscape, information about alternative binding routes, also relatively far away from the guess path, is going to be provided and the one associated to the minimum free energy determined. Structural features of metastable states can be extracted and, potentially, transition states between these relevant free energy minima can be recognized. In theory, besides energetics, insights about kinetics can be also achieved. Provided that the transition states are appropriately sampled and the related energy barriers accurately determined, and introducing some approximations for a reliable estimate of the pre-exponential factor, kinetic rates can be eventually computed.

3.2.4 Conclusions

In this study, we laid the ground for the estimation of the free energy profile associated to a protein-ligand binding process through computer simulation. Determining the FES for a complex reaction, such as a binding process, can be achieved by means of the path CVs. These CVs guide the sampling along, and around, a "guess path" that needs to be provided as input to the enhanced sampling engine.[18] This is essentially a frameset of the system captured at intermediate steps along the process under investigation. Herein, we constructed a guess path for Alprenolol binding to the β2-AR,[111] in order to use it for subsequent path CVs-based MetaD simulations. MetaD would allow calculating the desired FES. However, in order to built the needed putative pathway, some sort of indication about the process is necessary. To this end, we took advantage of plain MD simulations carried out in the D. E. research group for the spontaneous binding of Alprenolol to the β2-AR.[111] Basing on the available data, we constructed a MSM for the protein-ligand binding process, and determined the most relevant states. Notably, most of these corresponded to the ones suggested in the original paper. Once the relevant states were identified, they were used as template to build a guess path. Finally, we projected both the relevant states from the MSM and the plain MD trajectories into the path CVs space in order to validate our parameterization. The result was promising, with the relevant states located at low values of the z variable, and most of the plain MD trajectories, notably those with strikingly similar features, placed within the limit of z that could be affordable in the subsequent MetaD production phase.

3.3 TEST CASE 3: hDAAO

3.3.1 Introduction

In a typical drug discovery pipeline, after promising scaffolds are identified and selected in the hit identification phase, in the subsequent optimization step one aims at improving pharmacokinetic and pharmacodynamic properties. These include for instance absorption and distribution in the former case, and binding affinity in the latter. Traditionally, the affinity of a potential drug, generally speaking a ligand, towards a pharmacological target has been expressed in terms of K_d, the dissociation constant, or

IC_{50}, the drug concentration leading to the half-maximal inhibition of a biological activity. However, there has been a shift of perspective over the last decade, as the importance of kinetic quantities has become increasingly evident.[4–6] Specifically, we refer to k_{on} and k_{off}, the association and dissociation rate constants, respectively.[4] However, most of the focus has been directed to k_{off}, as this is directly related to drug residence time, which is expressed as the inverse of the kinetic parameter.[5,10] While the critical significance of binding affinity as an estimate of drug potency remains doubtless, there are cases in which integrating such information with residence time is more convenient. For instance, this is particularly true when the duration of the pharmacological effect plays a substantial role in *in vivo* efficacy.[5] In this view, it is clear why possibly achieving a rational optimization of the kinetic properties becomes extremely desired.

During the drug optimization phase, the classic framework would require synthesizing, or in the best scenario purchasing, analogs of the scaffold under study, and subsequently performing biological assays in order to estimate affinity and kinetic properties. Several experimental techniques, such as SPR,[126] NMR[127] and fluorescence methods,[128] have become established tools to characterize kinetic features and determine kinetic parameters.[129] However, being able to integrate or potentially replace completely this procedure with a computational approach would have the effect of improving considerably the efficiency, not to say the accessibility, of the optimization strategy. From the computational standpoint, this would require assessing the atomic-level determinants underlying binding and unbinding kinetics. The more appropriate approach would demand for MD simulations of the entire binding and unbinding processes. However, besides being not straightforward, the application of MD to this type of problems neither would be efficient. As already stressed out abundantly, most of the limitation lies inside the timescales problem. Undoubtedly, enhanced sampling methods[16] are currently the most effective and promising tools to sample, and thus to achieve information about, slow events. Nevertheless, by means of the current frameworks and considering that computational resources are typically limited, we can aim at managing one to a few ligands at the time. In addition, such procedures do not provide outcomes in a timely manner, as usually requested in an optimization phase.

Recently, a more practical and accessible methodology has been introduced to tackle the limitations described above.[21] In particular, the focus is on residence time, the inverse of the dissociation rate constant.[5] Instead of aiming at calculating absolute values for kinetic observables, the goal is being able to rank ligands according to their residence time.

Once this is achieved, ligands can be classified and distinguished in faster and slower in terms of unbinding simulation time.[21,130] Slower ligands would then be selected for further improvements. As discussed in detail in the *Theory* chapter, the methodology is based on scaled MD,[19,20] as it allows observing unbinding events at considerably more accessible computational costs.

Herein, we applied the ranking procedure to human D-amino acid oxidase (hDAAO). The flavoprotein hDAAO catalyzes the oxidative deamination of D-amino acids with excellent stereospecificity.[131] Firstly, the enzyme oxidizes the substrate with the concomitant reduction of a molecule of FAD, which acts as a cofactor; secondly, the produced imino acid is released into the solvent, where it non-enzymatically hydrolyzes into the corresponding α-keto acid and ammonia. A schematic representation of the mechanism is given in Figure 25.

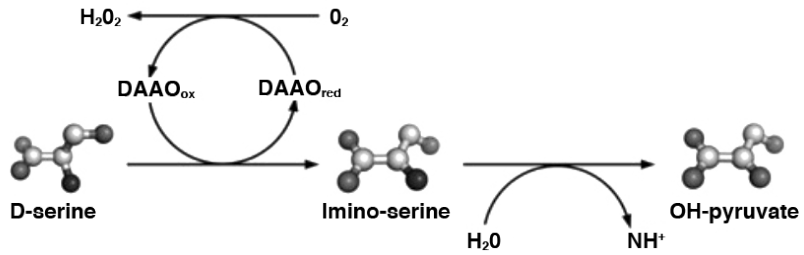

Figure 25. Mechanism of the D-serine oxidative deamination catalysed by hDAAO. Adapted from Ref.[131]

In humans, the enzyme is mainly present in liver and kidney, where it detoxifies D-amino acids from dietary and bacterial origin,[132] and in several regions of the brain, from cerebellum to frontal cortex, where it is involved in regulation of D-serine levels.[133–135] In the central neuronal system, the amino acid D-serine acts as a synaptic co-agonist at the NR1 subunit of the N-methyl-D-aspartate receptors (NMDARs), activation of which is associated with synaptic plasticity, learning and memory, and pain sensation.[136] From a pharmacological standpoint, D-serine dependent, aberrant NMDARs activity is involved in several neurological diseases. In particular, it has been highlighted that the lowered activity of the receptors is associated to psychiatric disorders, such as schizophrenia[137] and neuropathic pain.[138] The role of hDAAO in the brain was unknown until recently, when reciprocal correlation between D-serine and hDAAO concentration was demonstrated, suggesting the involvement of the enzyme in the metabolism of the neuromodulator.[133–135] Therefore, by increasing D-serine levels and thus activation of NMDARs, inhibition of

hDAAO has been proposed as a potential effective therapeutic approach. Since determination of the crystal structure in 2006, many efforts have been focused on development of potent inhibitors of the enzyme.[137]

Inspired by the result of a previous virtual screening campaign carried out in our laboratory, we selected among the reported compounds those that shared a chemical scaffold. Without any a priori knowledge of kinetic properties, we prioritized the compounds according to the computational time of unbinding from our simulations. Subsequently, kinetic data were produced and compared with the predicted values. The outcoming picture showed a good agreement between simulation and experiments. Moreover, from the analysis of our scaled MD trajectories, we were able to identify two major conformations for the loop located at the entrance of the hDAAO binding site.[139,140]

3.3.2 Methods

A first virtual screening campaign carried out in our lab, and aimed at identifying inhibitors of hDAAO, led to the purchase of 24 chemical compounds. Among these, 6 active compounds were identified in the μM regime. The goal was a competitive inhibition by targeting the protein binding pocket, where the oxidation of the substrate takes place through reduction of a molecule of the FAD cofactor.[131] The activity of the ligands identified ranged from about 10 to about 500 μM. Basing on these results, purchasable analogs of the compounds were searched by means of the SciFinder engine.[141] Thus, other 19 chemicals were gathered and activity assays performed. Scaffolds already reported in the literature were present among all of the selected compounds. However, no data were previously reported for the molecules considered in the virtual screening campaign. From the assays, no significant improvements were observed. Nevertheless, a group comprising 4 of the compounds was particularly interesting, as minor differences in chemical substituents were responsible for considerably different values of activity. Moreover, while IC_{50} values were available, no information about kinetic properties was provided. Inspired by all of these factors, we considered applying the ranking procedure reported in previous studies in order to prioritize these compounds in terms of unbinding simulation times.[21] One important aspect when applying the mentioned procedure is that, as a starting point for the scaled MD simulations, the structures of the ligands in complex with the protein are needed. In the literature, no crystal structure of hDAAO in complex with the scaffold

representing the selected compounds was reported. However, the crystal for a closely related small molecule was available. While the scaffold was different, the portion giving the core interactions was perfectly identical. Therefore, such crystal structure was employed as a template to position the ligands inside the binding pocket. As the scaffold without any substituent was previously reported in the literature,[142] we added it to the group of the considered compounds. Therefore, the 5 small molecules reported in Figure 26 were eventually considered.

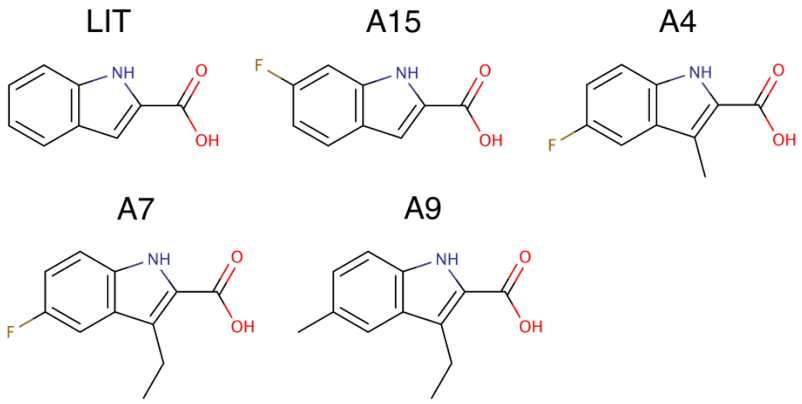

Figure 26. Structure of the compounds chosen for scaled MD simulations.

3.3.2.1 Simulation setup

As a first step, the complexes between ligands and hDAAO needed to be constructed. As already mentioned, no crystal structure was available in the Protein Data Bank (PDB) for the protein in complex with ligands possessing a common scaffold with the chosen compounds. Therefore, we took advantage of a closely related small molecule for which the PDB in complex with hDAAO was present (PDB ID: 3CUK, resolution 2.49 Å). The structure of the compound is reported in Figure 27B and compared with the common scaffold found for the 5 compounds. Focus on the binding site and the main interactions with the protein are also reported in Figure 27A. As shown in the Figure, the pyrrole-2-carboxylic acid moiety is responsible for all of the most significant interactions with the target. All of the chosen ligands maintained the identical moiety, thus we expected the same interactions to take place. Therefore, using 3CUK as a template, we created the complexes for our chemicals by placing them in hDAAO binding site with the Cartesian coordinates of the common moiety perfectly overlapping. The resulting superposition is shown in Figure 27C. Once the protein-ligand complexes were constructed, each system

was solvated in a cubic box with TIP3P water molecules.[88] In order to neutralize the overall charge, an adequate number of Na^+ ions was added.

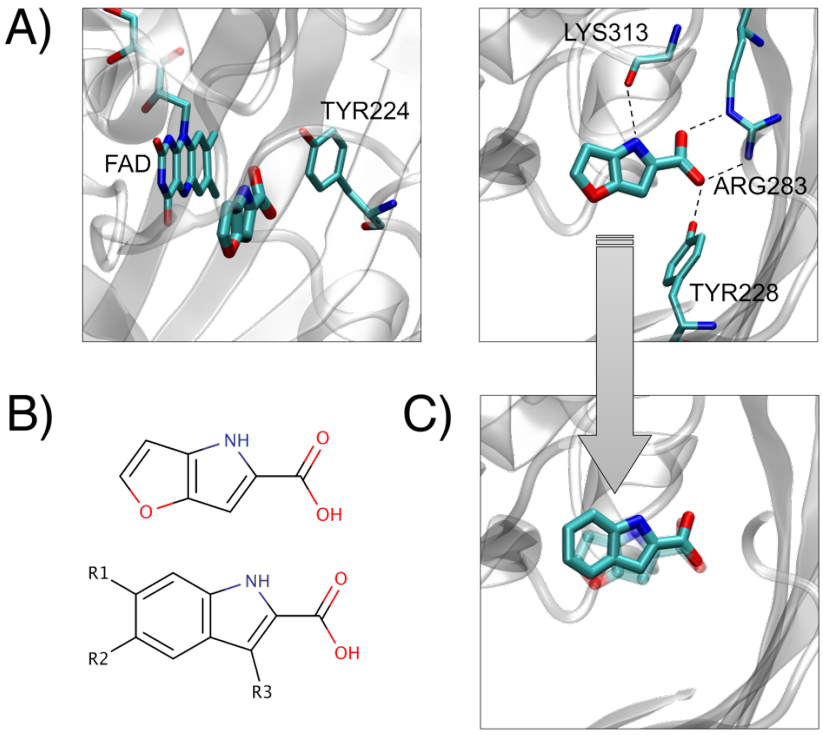

Figure 27. Creating the hDAAO-ligand complexes. A) Bound state for the crystal structure 3CUK. The ligand is packed between the flavin from FAD and the side-chain phenol from Tyr 224 (left panel). The pyrrol-2-carboxyl group gives rise to two hydrogen bonds, one with LYS313 and the other with TYR228, and a salt bridge, specifically with ARG283 (right panel). Such features are preserved in several PDBs for the protein in complex with different inhibitors. B) 2D structures for the ligand found in PDB 3CUK (top) and for the scaffold of the ligands selected for scaled MD simulations. The structures are shown for comparison. C) Superposition of the considered scaffold on the ligand of the 3CUK complex (shown as transparent).

The Amber ff99SB-ILDN[80,143] was employed for the protein, while the ligands were modeled according to the GAFF,[123] following the RESP procedure[124] to determine the charges. As already mentioned, substrate oxidation by hDAAO is concomitant with the reduction of a molecule of the FAD cofactor.[131] Notably, the cofactor flavin group is located at the very bottom of the substrate binding pocket. In all of the crystal complexes reported in the literature, the group is in close vicinity of ligands and it is oriented such as to give rise to a pi-pi stacking interaction with them. Therefore, we needed to include it in our system. Cofactors are usually composite molecules, typically comprising diverse chemical groups with different properties, as is the case for FAD. Moreover, due to their relatively large size, they are characterized by many degrees of freedom. Thus, deriving

proper, as to say reliable, parameters for such chemical species is far from trivial. For the present case, we took parameters for the FAD cofactor from the R.E.DD.B. database, where multiple conformations of each building blocks comprised in the molecules are considered in the systematic procedure followed to derive charges.[144]

According to the classic pipeline for system preparation, we first minimized and subsequently equilibrated each protein-ligand complex. In particular, position restraints of about 2.4 kcal/(mol Å^2) were first applied to all system heavy atoms and then to α-carbons and FAD and ligand heavy atoms in two subsequent steepest descent runs, 5000 step-long each. After this, during the equilibration phase, temperature was first increased in the NVT ensemble in three subsequent steps lasting 200 ps each. This was followed by a 400 ps-long run in NPT to relax the volume and, finally, a 500 ps run in the NVT ensemble were carried out under the same conditions employed for the subsequent production phase.

We took advantage of a GROMACS 4.6.1 version[93] appropriately modified in-house in order to perform scaled MD. As described extensively in the *Theory* chapter, under scaled MD conditions the PES of the system is scaled by a factor λ, where $0<\lambda<1$.[19] On the one hand, this has the effect of facilitating transitions along all of the degrees of freedom comprised in the system, in a similar way as simulating at high temperatures would do. On the other hand, a major drawback is that system stability is compromised as one applies lower, thus more aggressive, scaling factor values. More precisely, the secondary, and consequently the tertiary, structure of proteins tends to be disrupted. As a practical solution, the authors of the methodology applied weak position restraints in non-relevant regions of the considered protein. Specifically, values as low as 0.12 kcal/(mol Å^2) on backbone heavy atoms were sufficient to maintain protein structural features when a λ value of 0.4 was employed.[21,130] To choose a convenient scaling factor for our simulation, we performed some test runs on the fastest and slowest ligand at increasing λ values, such as 0.4, 0.45, 0.5 and 0.6. We note that values as low as 0.4 and 0.5 were applied in the previous works. In particular, in one of these cases,[130] it was shown that a significant improvement in ligand discrimination was achieved when repeating the procedure at the higher value of 0.5. While this was reasonable, as the effect on the PES responds to an exponential dependence, it translated into a considerably increased computational effort. It is worth noticing at this stage that no rule exists to determine the more appropriate scaling factor. While the method can be considered general as it affects the entire system, the choice of the λ factor is not. In our test runs, we were able to obtain a satisfactory

separation between the fastest and the slowest ligand, falling within the 100 ns range, when a scaling factor of 0.45 was employed. Therefore, such value was used to perform all of our scaled MD simulations. For what concerned the position restraints, taking the already mentioned crystal structure 3CUK as a reference, we selected those residues that were farther than a 6 Å distance from the ligand. All of the residues comprised in the loop located at the entrance of the binding site (residues 216 to 228)[139,140] were also excluded from the selection. Moreover, as the flavin ring from FAD was in close vicinity of the ligand, no restraints were set on it. As indicated in the previous works,[21,130] 0.12 kcal/(mol Å^2) weak restraints were applied to the backbone heavy atoms from the selected residues. Figure 28 gives a pictorial representation of the unrestrained regions of the protein.

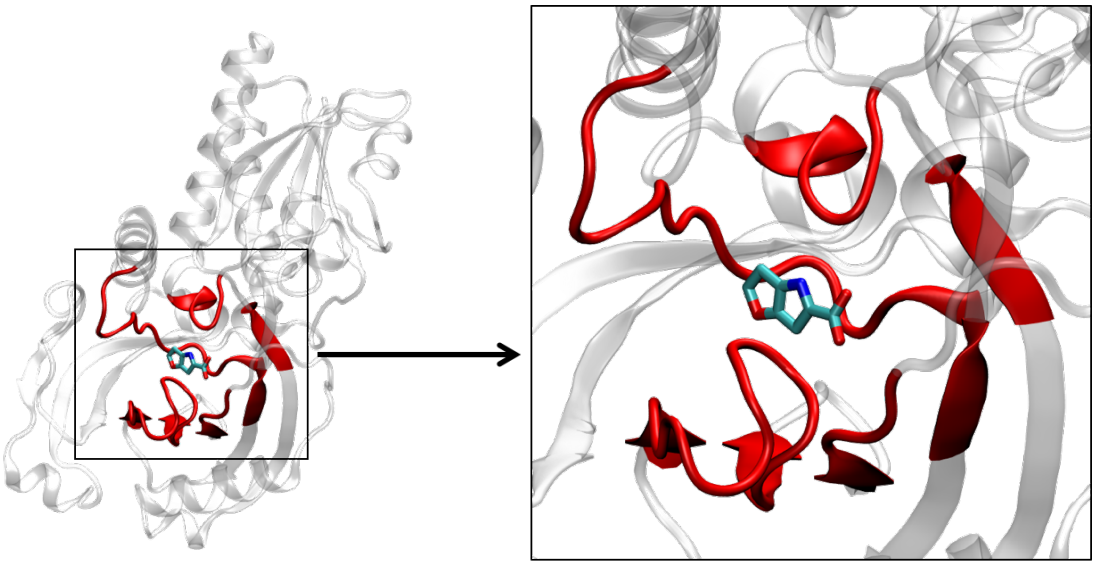

Figure 28. In order to preserve the overall protein structure, weak position restraints were applied to regions located far away from the binding site. In red colour, residues on which no restraints were set are highlighted.

Production runs were started from the final frame obtained in the last equilibration step, carried out in the NVT ensemble. For each simulation, initial velocities were randomized. In order to assess unbinding, the distance between ligand and pocket centers of mass was monitored as a function of the simulation time. When the distance reached a value of 30 Å, the run was stopped and the corresponding simulation time recorded. At such distances, ligand could be considered fully solvated and detached from the protein.[21] We considered such state as achieved unbinding. In case the protein-ligand complexes were not dissociated within 100 ns, the run was also terminated. As the major goal of this procedure was efficiency, we decided to contain the computational effort so as to be as low

as possible. Notably, no unbinding within the 100 ns took place more often for the slowest ligand only. For each of the 5 considered small molecules reported in Figure 26, we carried out 18 scaled MD simulations and recorder the computational unbinding times. We then took the mean of the values registered in order to rank the ligands according to their propensity to leave the binding site.

3.3.2.2 Enzymatic assays

Once the scaled MD simulations were performed, the compounds were purchased and experimental assays performed by our colleague Elena Rosini at The Protein Factory, Politecnico di Milano, ICMR-CNR, Università degli studi dell'Insubria.

To determine IC_{50} values, an Amplex Red-based assay has been utilized. This allowed measuring hDAAO product formation and thus determining the inhibitory activity for the compounds under investigation. hDAAO, horseradish peroxidase (HRP), FAD, as long as the considered compound were incubated for 30 minutes. After this pre-incubation period, D-serine and Amplex Red were added and the reaction proceeded for 30 additional minutes. A fluorescent product caused by hydrogen peroxide-dependent Amplex Red oxidation during hDAAO-catalyzed substrate turnover was measured in endpoint mode (excitation and emission wavelengths of 530 and 590 nm, respectively). The final concentration for the reactive components were as follows: 50 mM sodium phosphate, pH 7.4, 0.06 mg/mL human serum albumin, 7 nM His-hDAAO, 0.1 units/mL HRP, 4 μM FAD, 35 μM Amplex Red, 5mM D-serine, 0.8% (w/v) dimethyl sulfoxide (DMSO), 0-1.25 mM compound inhibitor. All enzymatic assays were conducted at room temperature in 96-well plate format. Data were fit to a standard, four parameters equation to determine curve top, bottom, concentration producing 50% inhibition (the IC_{50}) and hill slope.

Equilibrium binding constants were determined via absorption spectroscopy. The binding of the different small molecules was investigated by adding increasing concentrations of the ligand to a fixed amount of hDAAO (10 μM) in presence of 4 μM free FAD in 50 mM potassium phosphate buffer, pH equal to 7.5, and at 15 ºC. Absorption spectra were collected in the 300-800 nm range. The dissociation constant K_d, hereafter referred to as static K_d, was determined from the change in absorbance at about 492-494 nm in response of increasing ligand concentrations.

Association and dissociation kinetic constants k_{on} and k_{off} were calculated by means of stopped flow spectrophotometry. The dynamic dissociation constant, expressed as dynamic

K_d, was also determined from the ratio of the obtained values of the kinetic rates. Under the same experimental conditions as indicated in the previous paragraph, hDAAO (1.1 mg/mL) was rapidly mixed in a SFM-300 Biologic apparatus with a similar volume of increasing concentrations of inhibitor. The time course of spectral changes was recorder in the 300-700 nm wavelength range, and the absorbance versus time data sets were extrapolated at the wavelength identified by static titrations. Observed rate constants (k_{obs}) at increasing inhibitor concentrations were determined from the time courses by nonlinear regression using single exponential equations. The rate constants of inhibitor association and dissociation were determined by linear regression of the equation k_{obs} versus inhibitor concentration.

3.3.3 Results and discussion

We performed 18 scaled MD simulations for each one of the 5 ligands selected (Figure 26). Runs were stopped once unbinding took place or if no detaching of the compounds was observed within a computational time of 100 ns. As we already stressed, no information about kinetics was available at this point. Moreover, the considered compounds presented the same scaffold and only minor chemical substitutions. Notably, such substituent groups did not introduce the possibility to accomplish any additional hydrogen bond nor salt bridges. Thus, we applied the methodology to a particularly challenging framework. The resolution of the procedure was tested in a real-life situation, that is exploiting computer simulations in a prospective manner with a choice of compounds that reflects a likely picture one might encounter during the optimization phase. As far as we know, such a scenario was not considered before. The unbinding times that we obtained were recorded and, according to the procedure followed in previous works, the corresponding average, median, standard deviation and standard error values were calculated over the 18 runs performed for each ligand. The results are reported in Table 3.

As highlighted by the outcomes shown in the table, the ligands were characterized by considerably different average unbinding times, despite the small differences in their structures. In particular, ligand A9 was the fastest at leaving the binding site, while LIT was the slowest.

Ligand	A9	A4	A7	A15	LIT
Avg. (ns)	17.7	25.8	28.4	55.5	62.5
Median (ns)	18	20.8	25.9	50	51.2
St. Err.	3.2	4.3	4.2	8.1	6.4
St. Dev.	13.4	18.2	17.7	34.2	27.2

Table 3. Average, median, standard deviation and standard error obtained for the 5 ligands. The values were computed from the 18 scaled MD simulations performed for each ligand A9, A4, A7, A15 and LIT.

At a more comprehensive glance, we could distinguish two major regimes. Ligands A9, A4 and A7 could be grouped together in a fast regime, while ligands A15 and LIT in a slow regime. The separation was justified by an increase of more than 25 ns in terms of average unbinding time. Notably, while the first group is characterized by ligands possessing methyl or ethyl substituents on the pyrrol, this is not the case for ligands A15 and LIT. In other words, the hindrance caused by chemical groups on that side of the molecules, pointing towards the bottom of the binding pocket, is unfavorable for a tight binding of the active site, and causes ligands to detach faster. Notably, structure-activity relationship (SAR) carried out on scaffolds binding hDAAO reported in the literature suggested how the presence of bulkier substituents, such as methyl and ethyl groups, tended to reduce the activity of the compounds.[145,146] Thus, our results were in line with known behavior for hDAAO inhibitors. Moreover, the observations also confirmed that the initial complexes that we reconstructed for the 5 ligands, for which no crystal structure was available, were reasonable.

In our opinion, what was most informative was the qualitative interpretation of the unbinding simulations. Figure 29 shows the distribution of the computational unbinding times for each ligand. More than focusing on the values of average unbinding time, the aforementioned separation was what was clearest according to the Figure.

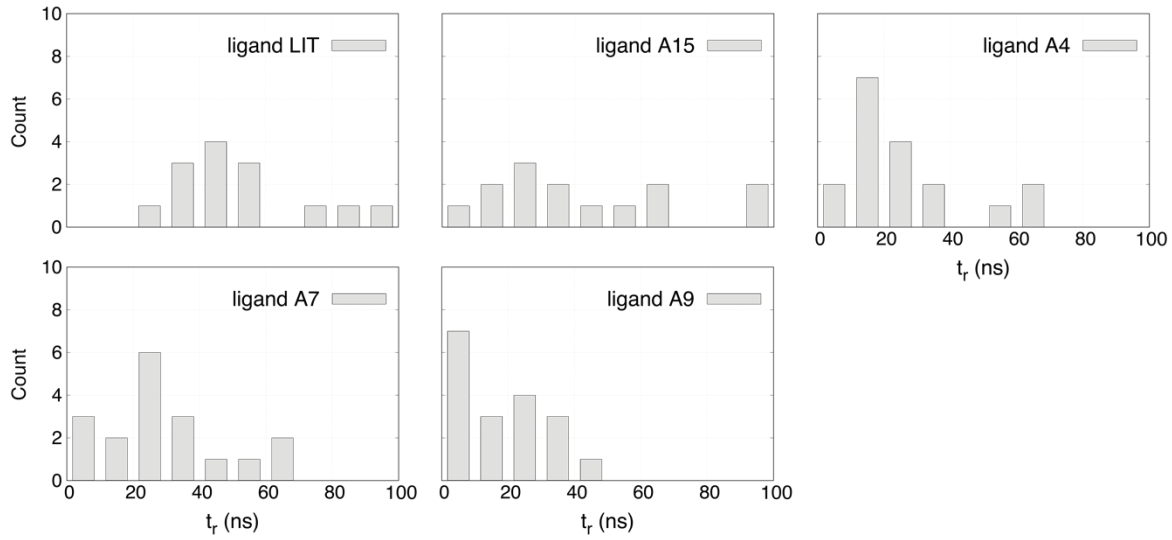

Figure 29. Distribution of the computational residence times. The counts for bins of 10 ns, from 0 to the maximum allowed value of 100 ns, are indicated on the y axis.

Going from ligand A9 to LIT, there was a shift on the distributions, with a gradual reduction of the amount of fast unbinding events and an increase of slow ones. According to this picture, one would have definitely gathered precious indications on ligands to further optimize or to discard in a drug optimization context.

Once the set of scaled MD simulations was completed, the compounds were purchased and assays to assess both activity and kinetic properties were carried out. The results are shown in Table 4. Unfortunately, despite ligand LIT was indicated as available for purchase, it was not possible to acquire it. Thus, we were able to carry out the experimental assays only for A9, A4, A7 and A15. Nevertheless, once purchase of the ligand LIT will be possible, we will integrate the results with the corresponding missing information.

Ligand	IC50 (μM)	Inactivation (%)	Static Kd (μM)	kon (μM/s)	koff (1/s)	Dynamic Kd (μM)
A9	250.81 ± 23.31	50	-	-	-	-
A4	10.63 ± 0.91	total	5.65 ± 0.73	0.15 ± 0.01	1.11 ± 0.08	7.4 ± 0.72
A7	14.5 ± 4.1	40	6.28 ± 0.71	0.12 ± 0.01	0.78 ± 0.08	6.5 ± 0.85
A15	5.38 ± 0.26	total	12.91 ± 2.35	0.23 ± 0.01	0.66 ± 0.06	2.9 ± 0.28

Table 4. Experimental data obtained from assays carried out on the 5 considered ligands. Uncertainties associated to the determined values are also indicated.

Surprisingly, notwithstanding a very subtle difference between A9 and A7, no binding was observed for the former. We adduce such behavior to an increase in steric hindrance

when replacing a fluorine with a methyl group. As we already outlined, the size of the hDAAO active site is relatively limited. Thus, the protein appears to be very sensible in terms of dimensions for the ligands to host in the binding pocket.

The trend of the k_{off} values determined through the experiments was in agreement with the one expected basing on the scaled MD runs. In particular, predicted increasing residence times from the simulations corresponded to decreasing experimental k_{off} values. The comparison is shown in Figure 30, where the correlation between simulations and experiments is reported.

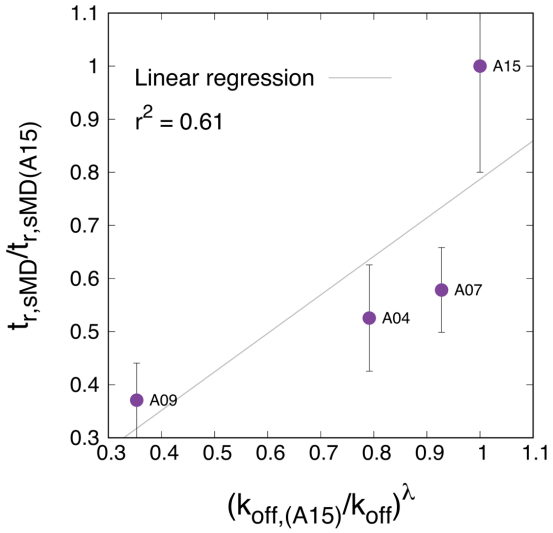

Figure 30. Computational versus experimental residence time. Values from both axes were normalized with respect to the slowest ligand A15; errors associated to the computational residence times, shown as error bars, were normalized as well, according to error propagation. Kinetic rates were subjected to exponential scaling according to the scaling factor adopted for performing the scaled MD simulations, that is 0.45. As no binding was reported for compound A9 in the assays, an arbitrary value was assigned; the choice was on one order of magnitude smaller than the dissociation rate constant determined for the slowest ligand.

It is worth noticing that no clear-cut separation between A7 and A15, corresponding to the one observed in the predicted unbinding time values, was present in the experiments. Nevertheless, A15 was distinguished as the slowest ligand among the ones tested, in agreement with what determined through the scaled MD runs. Therefore, in the perspective of prioritizing ligands in terms of their residence times, the qualitative interpretation of the results from the simulations was able to provide strikingly relevant indications, in line with the experimental assays. In light of this, ligand A9 would have been undoubtedly discarded while ligand A15 preserved, during a hypothetic drug optimization campaign.

A characteristic structural feature of hDAAO is the presence of a loop at the entrance of the binding site, comprising residues 216 to 228.[131] In the literature, it is typically referred to as the active-site lid.[139,140] Most of the hDAAO ligands posses an aromatic, double-ring moiety that tends to place in the active site, specifically stacked in between of the FAD flavin ring and the side chain phenol from Tyr 224.[145,147] Such configuration was already highlighted above and shown in Figure 27A. This stacked pi-pi scheme was observed in the bound state of several crystal complexes.[148,149] In such configuration, access to the binding site is hindered. This is clearly shown in Figure 31, where a surface representation of the protein is given and the region to access the binding pocket is indicated.

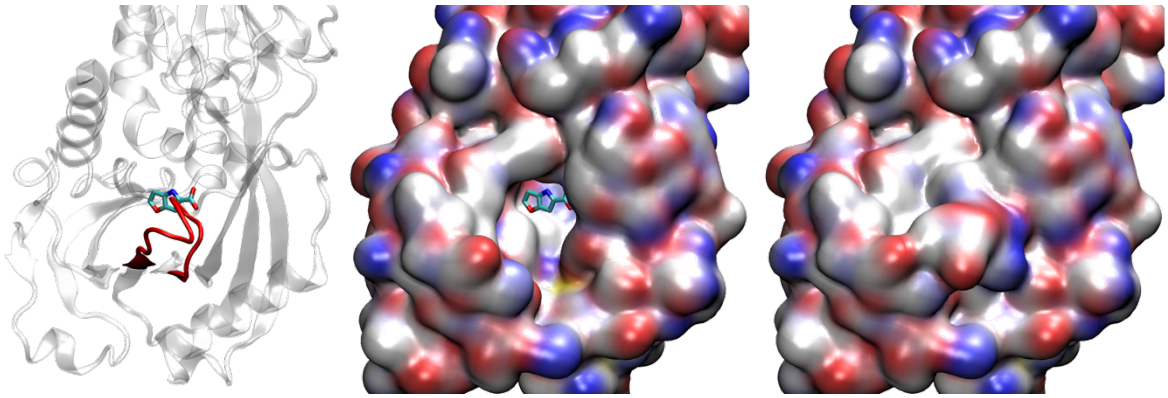

Figure 31. Focus on the active-site lid. The loop, comprising residues 216 to 228 (highlighted in red in the left hand panel), is located at the entrance of the binding site. In the central and right hand panels the protein is represent as vdW surface; in particular, the loop is intentionally not shown in the former. The comparison of the two surfaces highlights how the access to the ligand binding cavity is hindered by the presence of the active-site lid.

hDAAO binding pocket has a relatively small volume, and this feature is confirmed by the typical small size of known inhibitors. However, ligands of greater dimensions able to inhibit the enzyme were subsequently discovered. The crystal structures for complexes with these ligands bound showed a more open state of the active site lid.[139,149] This was necessarily related to the already mentioned small size of the pocket and to the binding pose of the inhibitors, that tend to direct one moiety towards the entrance of active site. Despites this behavior, no crystal structure with a completely open state for the active-site lid was reported up to this date. Moreover, while the possible involvement of this loop in ligand binding has been long proposed and argued,[139,140] no systematic characterization was carried out.

As we carried out unbinding simulations, we observed the active-site lid behavior in our scaled MD runs. As outline above, no position restraints were applied to the loop. Thus, this was able to rearrange and respond to ligand unbinding. We stress that, as already mentioned, transitions along all of the degrees of freedom of the system are significantly enhanced under scaled MD conditions. In other words, we had no means to assess that the observed active-site lid behavior reflected the real dynamics that would be recorded under plain MD conditions. This is particularly true as observing spontaneous unbinding (and also binding) in such a scenario, with a salt bridge, hydrogen bonds and pi-stacking stabilizing the bound state, and the binding site entrance closed by the loop, would likely require extremely long plain MD simulations. Nevertheless, we monitored the overall stability of the protein and the behavior of the loop by carrying out 3 plain MD simulations with no ligands in the binding site. In particular, as the active form of the enzyme is a homodimer,[131] we performed two 100 ns-long run for the monomer, and one 100 ns-long run for the dimer. The Root Mean Square Fluctuation (RMSF) on the α-carbons was calculated to observe mean fluctuations, and is shown in Figure 32.

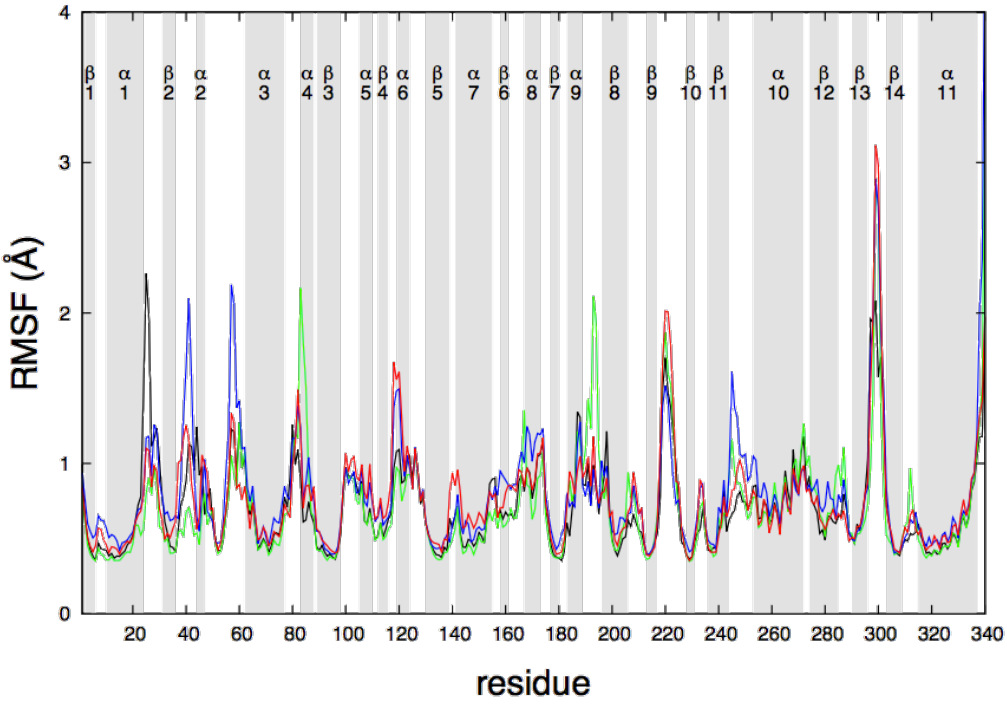

Figure 32. RMSF for the monomer in the plain MD runs. Four curves are shown, two corresponding to 100 ns-long simulations for the monomer, and two from the monomers present in the homodimer, for which one 100 ns long run was performed. The grey and white fingerprint in the background highlights structured, namely α-helices and β-sheets, and loop regions, respectively.

As it can be seen, beside high values for residues 295-305 that are comprised in a fully solvated loop located at a significant distance from the binding site, all of the runs demonstrated higher RMSF for residues 216-228. Despite no active-site lid opening/closure event was observed in the aggregated 400 ns of plain MD simulations for the monomer, this is undoubtedly an evidence of intrinsic flexibility for the active site loop.

Focusing on the active-site lid in the scaled MD runs during the unbinding process, we notice two major behaviors, to which we refer as pathway A and pathway B. In pathway A, Tyr 224 is fully solvated before the ligand is able to leave the active site. Whit the loop opened in such configuration, the entrance to the cavity is not hindered anymore and water molecules are able to access. This in turn facilitates the ligand detachment and unbinding takes place. Contrarily, when pathway B is followed, Tyr 224 points towards the binding pocket and is bent towards the base of the site. While the change with respect to the initial configuration is not as drastic as in pathway A, the cavity is also more accessible as a result. Thus, the ligand is able to squeeze through the available volume that connects with the bulk, once the stable interactions with protein residues are broken. To monitor the two possible cases, we calculated the distance between Tyr 224 and the binding site centers of mass as a function of the simulation time. Sample runs with unbinding in pathway A or pathway B are shown in Figure 33.

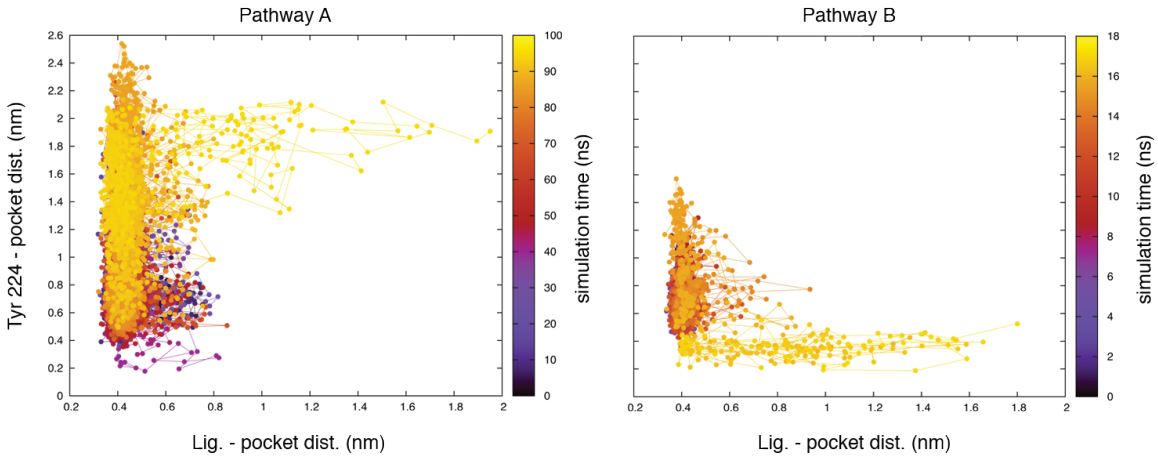

Figure 33. Sample pathways A (left panel) and B (right panel) during unbinding, respectively taken from scaled MD runs 15 and 16 on ligand A15. The ligand-pocket distance is shown on the x axis to assess unbinding. The values on the y axis was exploited to determined which of the two pathways was followed. In particular, a cut-off value of 10 Å (1 nm in the plots) was used to distinguish the two cases.

The frequency with which the two possible pathways were followed over the 18 scaled MD runs was registered for each one of the 5 ligands considered. The results are shown in Figure 34.

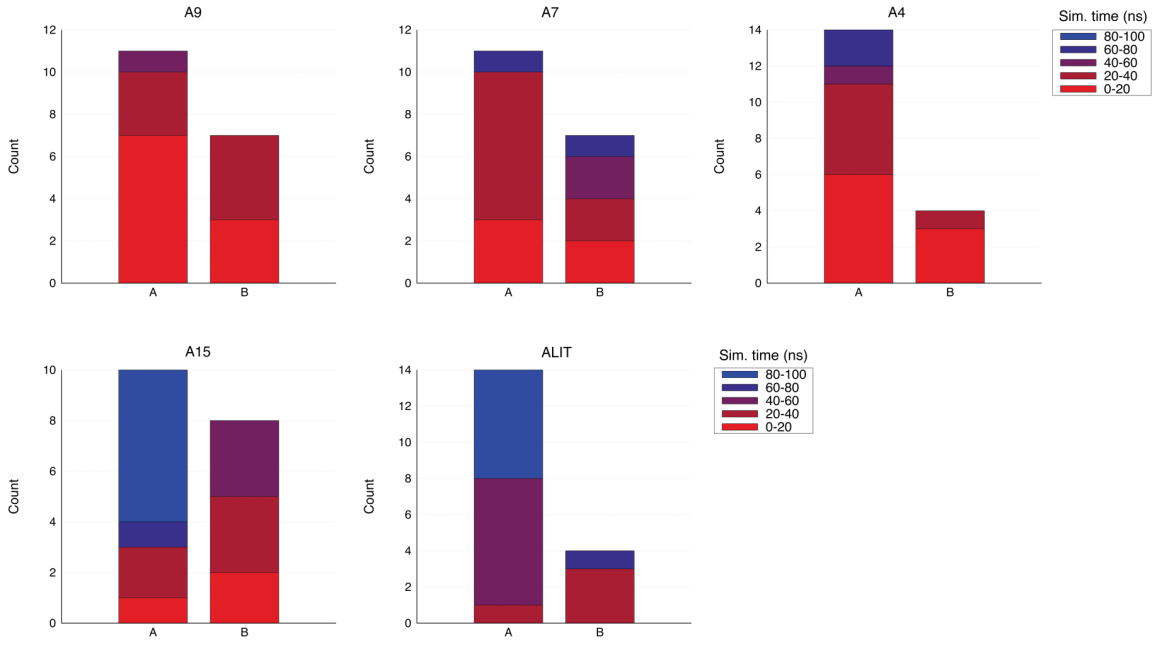

Figure 34. Classification of the unbinding events according to the path followed by the active-site lid, namely pathway A or pathway B. The corresponding bars are coloured basing on the simulation time of the unbinding events.

At a first glance, it is clear immediately that, in the vast majority of the cases, the pathway A was preferred. In other words, the opening of the lid was the limiting, necessary condition to allow the small molecules for jumping into the bulk. Notwithstanding the lower frequency with which the pathway B was followed, it is undoubtedly interesting observing that all of the ligands were able to access it. One relevant aspect arising is the different timescales related to the two paths. Notably, when pathway B was followed, unbinding was generally faster. Thus, while this route was less likely in our scaled MD simulations, it also required less time. Conversely, pathway A resulted as more likely, but longer times were involved to reach unbinding. We stress again that little can be concluded about the active-site lid behavior, as not enough statistics was gathered in the first place, and also because of the significantly smoothed potential energy surface due to the λ factor applied. Nevertheless, our observations were in line with a potential involvement of the loop in ligand binding/unbinding, and also suggested possible mechanisms. Certainly, assessing the real dynamics would require additional, and specifically devised, simulations

and possibly experiments. From such perspective, our observations might provide a useful starting point.

3.3.4 Conclusions

It is doubtless that residence time has gained significant relevance in drug discovery.[10] Thus, being able to improve kinetic properties during the drug optimization phase would be extremely desirable. For instance, devising inhibitors that give prolonged binding would extend the duration of the pharmacological effect, a particularly important point when this represents a large component of *in vivo* activity.[5] Within this framework, a computational approach based on enhanced sampling was recently developed to prioritize ligands according to their residence time.[21] Based on scaled MD, it relies on simulating several unbinding events for series of small molecules presenting a common scaffold. The average time necessary for unbinding is then calculated from the scaled MD runs for each ligand. The idea is to be able to discriminate between fast and slow binders, thus offering guidance in terms of kinetic properties during the optimization phase. In this study, we applied the procedure to a series of small molecules binding to hDAAO. The ligands, all possessing the same scaffold, were characterized by subtle modifications in term of substituent groups. Without any a priori knowledge about their residence time, we ranked these ligands according to the computational unbinding times. Subsequently, we performed assays to determine experimental off rates and the corresponding residence times. The correlation that we obtained was very satisfactory. Besides being able to distinguish between fast and slow binders, the ranking obtained through the computational procedure was confirmed by the experiments. Moreover, we investigated the scaled MD trajectories to possibly extract information about molecular features related to the unbinding process. Thus, we recognized two main conformations in which the active-site lid, located at the entrance of the binding pocket,[139,140] can be found during ligand unbinding. Despite no real quantification was achieved, this undoubtedly provided interesting indications that could be examined more thoroughly by subsequent studies.

3.4 AUTHOR CONTRIBUTION

The PhD candidate contributed substantially to all of the three applications presented herein. In particular, as for test case 1, namely N_{TAIL}, he performed the MetaD simulations and the analysis thereof. Additionally, he worked on the construction of the kinetic model proposed for the peptide. Concerning the second test case, that is the β2-AR complexed with the ligand Alprenolol, the candidate constructed the MSM starting from already available MD trajectories, performed the steered MD simulations, and built the guess path for the protein-ligand binding process. Finally, he carried out the scaled MD simulations on the third test case, hDAAO. All of the analysis of the produced trajectories, including the estimation of the computational unbinding times and the investigation about the active site lid behavior during ligand unbinding, were accomplished by the PhD candidate.

4. CONCLUSIVE REMARKS AND PERSPECTIVES

A comprehensive analysis of both structural and dynamical aspects pertaining to relevant biological systems demands for the inclusion of kinetics, alongside thermodynamics. In this dissertation, we explored possible strategies based on current, state-of-the-art computational techniques to achieve this end. We devised different protocols according to the specific biomolecular system and to the particular scientific question that we were interested in addressing. In all of the cases presented, despite recognized limitations, computational approaches proved to be effective and reliable tools for the considered purposes.

We showed that enhanced sampling, and in this specific case MetaD, can be exploited to characterize the highly heterogeneous configurational space accessible to an intrinsically disordered protein. The procedure was applied to N_{TAIL}, a test case IDP. In this particular context, force field represented indeed a more evident limitation than sampling itself. Although details about the timescales at which events take place are lost when employing enhanced sampling, we backtracked this information by constructing a kinetic model based on a binning strategy of the free energy. As a result, it was possible to estimate both thermodynamic and kinetic equilibrium properties of N_{TAIL}, achieving results in good agreement with available experimental data. The study clearly demonstrated how the proposed strategy is already approachable by means of current hardware and software architectures. The limiting steps are the reconstruction of a detailed and reliable FES and the determination of a diffusion matrix. Since the latter can be relatively easily obtained from multiple, short and not necessarily converged plain MD simulations, the former goal represents the major challenge. Undoubtedly, achieving a comprehensive and statistically relevant exploration of a configurational space is a more general matter in structural biology, that extends beyond the specific problem addressed in this dissertation. We dealt here with a single solute molecule possessing a highly heterogeneous conformational space. While systems of similar molecular size and complexity can be already addressed, even more challenging scenarios could be considered in the future, such as protein-protein or protein-ligand association. Therefore, as long as the FES related to specific processes of interest can be accurately reconstructed, the proposed procedure holds great potential in elucidating the underlying kinetics.

Unveiling the molecular features associated with protein-ligand binding would provide a striking support to the optimization of potential, novel drug-like molecules. Unfortunately, from a computational standpoint, achieving such detailed description with statistical significance is extremely onerous. In this context, we aimed at reconstructing the free energy surface for the binding process of the well-known inhibitor Alprenolol to the β2-AR. We constructed a Markov State Model taking advantage of long-timescale plain MD simulations to spot the relevant states along the binding route. MSM are typically constructed from plain MD trajectories, through count of the transitions between the different microstates in which the system can be found Thus, the model allows identifying the timescales associated with such transitions and predicting longer timescales kinetics. Notwithstanding the possibility of aggregating multiple, shorter, independent trajectories, that has increased significantly the appealing of such method, the demand in terms of sampling is still remarkable. Indeed, gathering appropriate statistics for each microstate, which is an essential requirement for the estimation of a reliable transition probability matrix, is currently the major challenge. In the case we faced herein, tens of microseconds were necessary to capture few spontaneous binding events. Notably, no unbinding nor rebinding were observed, making the information even more limited. Despites providing precious indications about the molecular features involved, these were not statistically meaningful. As such, we employed the data to build a non-reversible MSM from which determining the relevant states along the binding route. Basing on this information, we were able to reconstruct a putative pathway going from the protein surface to the orthosteric binding site, that will be subsequently exploited to perform path CV-based MetaD in order to determine the free energy surface associated with the process. In particular, energy barriers between metastable states could be ultimately determined and kinetic rates computed, achieving a level of characterization that is typically not accessible to experiments. While the proposed procedure can be successful, it is extremely case-dependent and applicable to one to few systems, considering the effort involved. Nevertheless, in light of the advances in computer power observed during the recent years, it is reasonable to expect that gathering sufficient statistics for complex processes via plain MD will be increasingly feasible. This, in turn, will allow constructing reliable MSM from which extracting kinetic information directly. Furthermore, recent theoretical advances are demonstrating the possibility of exploiting enhanced methods as a source of sampling for MSM construction, considerably accelerating the possibility of retrieving kinetics from molecular simulations.

Real life scenarios in drug discovery and development programs demand for practical, effective and relatively fast solutions, possibly applicable to a bunch of drug-like molecules. Both academia and mostly industry are gradually including protocols to evaluate kinetic parameters in the ligand optimization stages. However, from a computational standpoint, due to the complex, system-specific setup and the significant computational effort involved, achieving a detailed characterization of molecular determinants and the estimation of kinetic parameters in a routinely manner is not feasible yet. Nevertheless, strategies are emerging aiming at fulfilling such demand. We applied a recent procedure based on prioritization of ligands according to their relative residence times determined through unbinding simulations carried out via scaled MD. Specifically, we considered a series of congeneric ligand of hDAAO bearing subtle structural differences. Without any a priori information about their kinetic profiles, we showed that the obtained prioritization reflected the subsequently determined experimental kinetic data. The present application along with previous studies are demonstrating that the strategy is able to rank ligands in agreement with experiments. Assessing this consensus represents a crucial, necessary step towards the utilization of these simulations as a reliable, independent tool. From a drug discovery standpoint, the ability of discriminating faster and slower ligands is undoubtedly of great interest. Being able to prioritize those ligands possessing the desired kinetic features for further optimization would be of remarkable support. Despite being a promising picture, the computational effort involved is still not negligible. For instance, in the present case, where five ligands were examined, the protocol led to production of an amount of trajectories in the order of hundreds of nanoseconds. However, as already stressed, few to no binding events would be observed via plain MD for a single ligand within such simulation lengths. While the strategy is already emerging as a promising tool, the possibility of handling larger sets of compounds will express its real potential in the drug discovery pipeline. Future developments of current hardware and software architectures will be the major determinants towards this goal. In a more focused perspective, more gentle scaling factors could be also explored on representative ligands. Once an appropriate space is envisaged on which attempting a reweighting procedure, the underlying FES can be reconstructed and precious insights gained into the energetics and kinetics associated with the binding event. Notably, the latter scheme can find broader applications besides protein-ligand binding, including conformational changes and protein-protein association.

As a final remark, we can distinguish two main pictures arising from our studies. On the one side, more demanding computational protocols can be applied when particularly challenging scenarios need to be tackled, and specific questions need to be addressed. Although this does not configure well within routinely applicable procedures, it nevertheless represents a precious resource. On the other side, real life challenges require juggling multiple ligands and possibly multiple pharmacological targets. Therefore, computational chemistry strategies that are relatively easy to apply, and that involve contained computational resources, are starting to gain attention. Notwithstanding a possibly questionable accuracy in single, specific cases, an overall effectiveness when dealing with large numbers and aiming at a gross selection would nevertheless be a significant achievement.

REFERENCES

1. Haynie, D. T. *Biological Thermodynamics*. (Cambridge University Press, 2008). doi:DOI: 10.1017/CBO9780511802690
2. Glaser, R. in *Biophysics: An Introduction* (ed. Glaser, R.) 333–375 (Springer Berlin Heidelberg, 2012). doi:10.1007/978-3-642-25212-9_5
3. Mullard, A. Parsing clinical success rates. *Nat. Rev. Drug Discov.* **15,** 447 (2016).
4. 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).
5. 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).
6. Swinney, D. C. The role of binding kinetics in therapeutically useful drug action. *Curr. Opin. Drug Discov. Devel.* **12,** 31–39 (2009).
7. Keighley, W. The need for high throughput kinetics early in the drug discovery process. *Drug Discov World* **12,** 39–45 (2011).
8. Hasenhuetl, P. S. *et al.* Ligand selectivity among the dopamine and serotonin transporters specified by the forward binding reaction. *Mol. Pharmacol.* **88,** 12–18 (2015).
9. Kruse, A. C. *et al.* Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **482,** 552–556 (2012).
10. Copeland, R. A. The drug-target residence time model: a 10-year retrospective. *Nat. Rev. Drug Discov.* **15,** 87–95 (2016).
11. Vauquelin, G., Bostoen, S., Vanderheyden, P. & Seeman, P. Clozapine, atypical antipsychotics, and the benefits of fast-off D2 dopamine receptor antagonism. *Naunyn. Schmiedebergs. Arch. Pharmacol.* **385,** 337–372 (2012).
12. Katoh, E. *et al.* A solution NMR study of the binding kinetics and the internal dynamics of an HIV-1 protease-substrate complex. *Protein Sci.* **12,** 1376–1385 (2003).
13. Schneider, R. *et al.* Visualizing the Molecular Recognition Trajectory of an Intrinsically Disordered Protein Using Multinuclear Relaxation Dispersion NMR. *J. Am. Chem. Soc.* **137,** 1220–1229 (2015).
14. Aristotelous, T. *et al.* Discovery of β2 adrenergic receptor ligands using biosensor fragment screening of tagged wild-type receptor. *ACS Med. Chem. Lett.* **4,** 1005–1010 (2013).
15. Uversky, V. N. Intrinsically disordered proteins from A to Z. *Int. J. Biochem. Cell Biol.* **43,** 1090–1103 (2011).
16. Abrams, C. & Bussi, G. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy* **16,** 163–199 (2013).
17. Pande, V. S., Beauchamp, K. & Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52,** 99–105 (2010).
18. Branduardi, D., Gervasio, F. L. & Parrinello, M. From A to B in free energy space. *J Chem Phys* **126,** 54103 (2007).
19. 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).
20. Tsujishita, 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).
21. Mollica, L. *et al.* Kinetics of protein-ligand unbinding via smoothed potential

molecular dynamics simulations. *Sci Rep* **5,** 11539 (2015).
22. Zhou, H.-X. Rate theories for biologists. *Q. Rev. Biophys.* **43,** 219–293 (2010).
23. Pollak, E., Grabert, H. & Hänggi, P. Theory of activated rate processes for arbitrary frequency dependent friction: Solution of the turnover problem. *J Chem Phys* **91,** 4073–4087 (1989).
24. 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).
25. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc Natl Acad Sci U S A* **99,** 12562–12566 (2002).
26. Marinelli, F., Pietrucci, F., Laio, A. & Piana, S. A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. *PLoS Comput Biol* **5,** e1000452 (2009).
27. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23,** 187–199 (1977).
28. Isralewitz, B., Gao, M. & Schulten, K. Steered molecular dynamics and mechanical functions of proteins. *Curr Opin Struct Biol* **11,** 224–230 (2001).
29. Grubmuller, H., Heymann, B. & Tavan, P. Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. *Science (80-.).* **271,** 997–999 (1996).
30. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314,** 141–151 (1999).
31. Barducci, A., Bonomi, M. & Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1,** 826–843 (2011).
32. Barducci, A., Bussi, G. & Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **100,** 20603 (2008).
33. Piana, S. & Laio, A. A Bias-Exchange Approach to Protein Folding. *J. Phys. Chem. B* **111,** 4553–4559 (2007).
34. Nadler, W. & Hansmann, U. H. E. Optimized Explicit-Solvent Replica Exchange Molecular Dynamics from Scratch. *J. Phys. Chem. B* **112,** 10386–10387 (2008).
35. Bonomi, M. & Parrinello, M. Enhanced Sampling in the Well-Tempered Ensemble. *Phys. Rev. Lett.* **104,** 190601 (2010).
36. Deighan, M., Bonomi, M. & Pfaendtner, J. Efficient Simulation of Explicitly Solvated Proteins in the Well-Tempered Ensemble. *J. Chem. Theory Comput.* **8,** 2189–2192 (2012).
37. Bowman, G. R. in *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* 7–22 (Springer, 2014).
38. Prinz, J. H. *et al.* Markov models of molecular kinetics: generation and validation. *J Chem Phys* **134,** 174105 (2011).
39. Senne, M., Trendelkamp-Schroer, B., Mey, A. S. J. S., Schütte, C. & Noé, F. EMMA: a software package for Markov model building and analysis. *J. Chem. Theory Comput.* **8,** 2223–2238 (2012).
40. Plattner, N., Doerr, S., De Fabritiis, G. & Noé, F. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* **9,** 1005 (2017).
41. Pinamonti, G. *et al.* Predicting the Kinetics of RNA Oligonucleotides Using Markov State Models. *J. Chem. Theory Comput.* **13,** 926–934 (2017).
42. Pérez-Hernández, G., Paul, F., Giorgino, T., Fabritiis, G. De & Noé, F. Identification of slow molecular order parameters for Markov model construction. *J Chem Phys* **139,** 15102 (2013).
43. MacQueen, J. Some methods for classification and analysis of multivariate

observations. in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* **1,** 281–297 (Oakland, CA, USA., 1967).
44. Steinhaus, H. Sur la division des corp materiels en parties. *Bull. Acad. Pol. Sci* **1,** 801 (1956).
45. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **2,** 559–572 (1901).
46. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24,** 417 (1933).
47. Noé, F. & Clementi, C. Kinetic distance and kinetic maps from molecular dynamics simulation. *J. Chem. Theory Comput.* **11,** 5002–5011 (2015).
48. Schwantes, C. R. & Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **9,** 2000–2009 (2013).
49. Schütte, C., Fischer, A., Huisinga, W. & Deuflhard, P. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.* **151,** 146–168 (1999).
50. Deuflhard, P., Huisinga, W., Fischer, A. & Schütte, C. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra Appl.* **315,** 39–59 (2000).
51. Deuflhard, P. & Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Appl.* **398,** 161–184 (2005).
52. Noé, F., Horenko, I., Schütte, C. & Smith, J. C. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J Chem Phys* **126,** 04B617 (2007).
53. Voter, A. F. in *Radiation Effects in Solids* (eds. Sickafus, K. E., Kotomin, E. A. & Uberuaga, B. P.) 1–23 (Springer Netherlands, 2007). doi:10.1007/978-1-4020-5295-8_1
54. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* **293,** 321–331 (1999).
55. Dunker, A. K. *et al.* Intrinsically disordered protein. *J Mol Graph Model* **19,** 26–59 (2001).
56. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6,** 197–208 (2005).
57. Dunker, A. K., Silman, I., Uversky, V. N. & Sussman, J. L. Function and structure of inherently disordered proteins. *Curr Opin Struct Biol* **18,** 756–764 (2008).
58. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z. & Dunker, A. K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* **323,** 573–584 (2002).
59. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* **37,** 215–246 (2008).
60. Felli, I. C. & Pierattelli, R. Recent progress in NMR spectroscopy: toward the study of intrinsically disordered proteins of increasing size and complexity. *IUBMB Life* **64,** 473–481 (2012).
61. Kikhney, A. G. & Svergun, D. I. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett* **589,** 2570–2577 (2015).
62. Schuler, B., Müller-Späth, S., Soranno, A. & Nettels, D. Intrinsically Disordered Protein Analysis: Vol. 2, Methods and Experimental Tools. (2012).
63. Abyzov, A. *et al.* Identification of Dynamic Modes in an Intrinsically Disordered

Protein Using Temperature-Dependent NMR Relaxation. *J. Am. Chem. Soc.* **138,** 6240–6251 (2016).
64. Do, T. N., Choy, W.-Y. & Karttunen, M. Accelerating the Conformational Sampling of Intrinsically Disordered Proteins. *J. Chem. Theory Comput.* **10,** 5081–5094 (2014).
65. Wang, Y. *et al.* Multiscaled exploration of coupled folding and binding of an intrinsically disordered molecular recognition element in measles virus nucleoprotein. *Proc Natl Acad Sci U S A* **110,** E3743-52 (2013).
66. Salvi, N., Abyzov, A. & Blackledge, M. Multi-Timescale Dynamics in Intrinsically Disordered Proteins from NMR Relaxation and Molecular Simulation. *J. Phys. Chem. Lett.* **7,** 2483–2489 (2016).
67. Stanley, N., Esteban-Martín, S. & De Fabritiis, G. Progress in studying intrinsically disordered proteins with atomistic simulations. *Prog. Biophys. Mol. Biol.* **119,** 47–52 (2015).
68. Voelz, V. A., Bowman, G. R., Beauchamp, K. & Pande, V. S. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J Am Chem Soc* **132,** 1526–1528 (2010).
69. Hansmann, U. H. E. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.* **281,** 140–150 (1997).
70. Bussi, G., Gervasio, F. L., Laio, A. & Parrinello, M. Free-Energy Landscape for β Hairpin Folding from Combined Parallel Tempering and Metadynamics. *J. Am. Chem. Soc.* **128,** 13435–13441 (2006).
71. Bellucci, L., Bussi, G., Di Felice, R. & Corni, S. Fibrillation-prone conformations of the amyloid-beta-42 peptide at the gold/water interface. *Nanoscale* **9,** 2279–2290 (2017).
72. Rossetti, G. *et al.* Conformational ensemble of human alpha-synuclein physiological form predicted by molecular simulations. *Phys Chem Chem Phys* **18,** 5702–5706 (2016).
73. Fisher, C. K. & Stultz, C. M. Constructing ensembles for intrinsically disordered proteins. *Curr Opin Struct Biol* **21,** 426–431 (2011).
74. Robustelli, P., Kohlhoff, K., Cavalli, A. & Vendruscolo, M. Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. *Structure* **18,** 923–933 (2010).
75. Camilloni, C., Cavalli, A. & Vendruscolo, M. Assessment of the Use of NMR Chemical Shifts as Replica-Averaged Structural Restraints in Molecular Dynamics Simulations to Characterize the Dynamics of Proteins. *J. Phys. Chem. B* **117,** 1838–1843 (2013).
76. Granata, D. *et al.* The inverted free energy landscape of an intrinsically disordered peptide by simulations and experiments. *Sci Rep* **5,** 15449 (2015).
77. Skiadopoulos, M. H. *et al.* Sendai virus, a murine parainfluenza virus type 1, replicates to a level similar to human PIV1 in the upper and lower respiratory tract of African green monkeys and chimpanzees. *Virology* **297,** 153–160 (2002).
78. Jensen, M. R. *et al.* Quantitative Conformational Analysis of Partially Folded Proteins from Residual Dipolar Couplings: Application to the Molecular Recognition Element of Sendai Virus Nucleoprotein. *J. Am. Chem. Soc.* **130,** 8055–8061 (2008).
79. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophys J* **100,** L47-9 (2011).
80. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of

improved protein backbone parameters. *Proteins* **65,** 712–725 (2006).
81. Best, R. B. & Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix−Coil Transition of Polypeptides. *J. Phys. Chem. B* **113,** 9004–9015 (2009).
82. Best, R. B. & Mittal, J. Protein Simulations with an Optimized Water Model: Cooperative Helix Formation and Temperature-Induced Unfolded State Collapse. *J. Phys. Chem. B* **114,** 14916–14923 (2010).
83. Abascal, J. L. & Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *J Chem Phys* **123,** 234505 (2005).
84. Piana, S., Donchev, A. G., Robustelli, P. & Shaw, D. E. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J Phys Chem B* **119,** 5113–5123 (2015).
85. Ahalawat, N., Arora, S. & Murarka, R. K. Structural Ensemble of CD4 Cytoplasmic Tail (402-419) Reveals a Nearly Flat Free-Energy Landscape with Local alpha-Helical Order in Aqueous Solution. *J Phys Chem B* **119,** 11229–11242 (2015).
86. Cino, E. A., Choy, W. Y. & Karttunen, M. Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations. *J Chem Theory Comput* **8,** 2725–2740 (2012).
87. Lindorff-Larsen, K. *et al.* Systematic validation of protein force fields against experimental data. *PLoS One* **7,** e32131 (2012).
88. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79,** 926–935 (1983).
89. Joung, I. S. & Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **112,** 9020–9041 (2008).
90. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J Chem Phys* **126,** 14101 (2007).
91. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52,** 7182–7190 (1981).
92. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* **98,** 10089–10092 (1993).
93. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **4,** 435–447 (2008).
94. 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).
95. Pietrucci, F. & Laio, A. A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures: Application to SH3 and GB1. *J. Chem. Theory Comput.* **5,** 2197–2201 (2009).
96. Vymětal, J. & Vondrášek, J. Gyration- and Inertia-Tensor-Based Collective Coordinates for Metadynamics. Application on the Conformational Behavior of Polyalanine Peptides and Trp-Cage Folding. *J. Phys. Chem. A* **115,** 11455–11465 (2011).
97. Pietrucci, F., Mollica, L. & Blackledge, M. Mapping the Native Conformational Ensemble of Proteins from a Combination of Simulations and Experiments: New Insight into the src-SH3 Domain. *J. Phys. Chem. Lett.* **4,** 1943–1948 (2013).
98. Mollica, L. *et al.* Atomic-Resolution Structural Dynamics in Crystalline Proteins from NMR and Molecular Simulation. *J Phys Chem Lett* **3,** 3657–3662 (2012).
99. Shen, Y. & Bax, A. SPARTA+: a modest improvement in empirical NMR chemical

shift prediction by means of an artificial neural network. *J Biomol NMR* **48,** 13–22 (2010).
100. Stein, W. A. Sage Mathematics Software, version 5.4. (2012).
101. Li, D.-W. & Brüschweiler, R. Certification of Molecular Dynamics Trajectories with NMR Chemical Shifts. *J. Phys. Chem. Lett.* **1,** 246–248 (2010).
102. Palazzesi, F., Barducci, A., Tollinger, M. & Parrinello, M. The allosteric communication pathways in KIX domain of CBP. *Proc. Natl. Acad. Sci.* **110,** 14237–14242 (2013).
103. Han, M., Xu, J., Ren, Y. & Li, J. Simulation of coupled folding and binding of an intrinsically disordered protein in explicit solvent with metadynamics. *J. Mol. Graph. Model.* **68,** 114–127 (2016).
104. Park, H.-S. & Jun, C.-H. A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.* **36,** 3336–3341 (2009).
105. Konrat, R. NMR contributions to structural dynamics studies of intrinsically disordered proteins. *J Magn Reson* **241,** 74–85 (2014).
106. Uversky, V. N. Paradoxes and wonders of intrinsic disorder: Complexity of simplicity. *Intrinsically Disord Proteins* **4,** e1135015 (2016).
107. Palazzesi, F., Prakash, M. K., Bonomi, M. & Barducci, A. Accuracy of Current All-Atom Force-Fields in Modeling Protein Disordered States. *J. Chem. Theory Comput.* **11,** 2–7 (2015).
108. Houben, K., Marion, D., Tarbouriech, N., Ruigrok, R. W. H. & Blanchard, L. Interaction of the C-terminal domains of sendai virus N and P proteins: comparison of polymerase-nucleocapsid interactions within the paramyxovirus family. *J. Virol.* **81,** 6807–6816 (2007).
109. Higo, J., Nishimura, Y. & Nakamura, H. A free-energy landscape for coupled folding and binding of an intrinsically disordered protein in explicit solvent from detailed all-atom computations. *J Am Chem Soc* **133,** 10448–10458 (2011).
110. Joshi, P. & Vendruscolo, M. in *Intrinsically Disordered Proteins Studied by NMR Spectroscopy* 383–400 (2015). doi:10.1007/978-3-319-20164-1_13
111. Dror, R. O. *et al.* Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc Natl Acad Sci U S A* **108,** 13118–13123 (2011).
112. Buch, I., Giorgino, T. & De Fabritiis, G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc Natl Acad Sci U S A* **108,** 10184–10189 (2011).
113. Huang, D. & Caflisch, A. The free energy landscape of small molecule unbinding. *PLoS Comput Biol* **7,** e1002002 (2011).
114. Shan, Y. *et al.* How does a drug molecule find its target binding site? *J Am Chem Soc* **133,** 9181–9183 (2011).
115. Scherer, M. K. *et al.* PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **11,** 5525–5542 (2015).
116. Plattner, N. & Noé, F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat. Commun.* **6,** 7653 (2015).
117. Xu, D. & Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Ann. Data Sci.* **2,** 165–193 (2015).
118. Weinan, E. & Vanden-Eijnden, E. Towards a Theory of Transition Paths. *J. Stat. Phys.* **123,** (2006).
119. Metzner, P., Schütte, C. & Vanden-Eijnden, E. Transition path theory for Markov jump processes. *Multiscale Model. Simul.* **7,** 1192–1219 (2009).
120. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of

Molecular Dynamics Trajectories. *Biophys J* **109,** 1528–1532
121. Case, D. A. *et al.* Amber 14. (2014).
122. Dickson, C. J. *et al.* Lipid14: the amber lipid force field. *J. Chem. Theory Comput.* **10,** 865–879 (2014).
123. 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).
124. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **97,** 10269–10280 (1993).
125. Catmull, E. & Rom, R. in *Computer aided geometric design* 317–326 (Elsevier, 1974).
126. Patching, S. G. Surface plasmon resonance spectroscopy for characterisation of membrane protein-ligand interactions and its potential for drug discovery. *Biochim. Biophys. Acta* **1838,** 43–55 (2014).
127. Millet, O., Bernadó, P., Garcia, J., Rizo, J. & Pons, M. NMR measurement of the off rate from the first calcium-binding site of the synaptotagmin I C2A domain. *FEBS Lett.* **516,** 93–96 (2002).
128. Sridharan, R., Zuber, J., Connelly, S. M., Mathew, E. & Dumont, M. E. Fluorescent approaches for understanding interactions of ligands with G protein coupled receptors. *Biochim. Biophys. Acta* **1838,** 15–33 (2014).
129. 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).
130. 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).
131. Sacchi, S., Caldinelli, L., Cappelletti, P., Pollegioni, L. & Molla, G. Structure–function relationships in human d-amino acid oxidase. *Amino Acids* **43,** 1833–1850 (2012).
132. Konno, R., Niwa, A. & Yasumura, Y. Intestinal bacterial origin of D-alanine in urine of mutant mice lacking D-amino-acid oxidase. *Biochem. J.* **268,** 263–5 (1990).
133. Labrie, V., Clapcote, S. J. & Roder, J. C. Mutant mice with reduced NMDA-NR1 glycine affinity or lack of d-amino acid oxidase function exhibit altered anxiety-like behaviors. *Pharmacol. Biochem. Behav.* **91,** 610–620 (2009).
134. Sacchi, S. *et al.* pLG72 modulates intracellular D-serine levels through its interaction with D-amino acid oxidase: Effect on schizophrenia susceptibility. *J. Biol. Chem.* **283,** 22244–22256 (2008).
135. Wolosker, H. Serine racemase and the serine shuttle between neurons and astrocytes. *Biochimica et Biophysica Acta - Proteins and Proteomics* **1814,** 1558–1566 (2011).
136. Panatier, A. *et al.* Glia-derived D-serine controls NMDA receptor activity and synaptic memory. *Cell* **125,** 775–784 (2006).
137. Sacchi, S., Rosini, E., Pollegioni, L. & Molla, G. D-amino acid oxidase inhibitors as a novel class of drugs for schizophrenia therapy. *Curr. Pharm. Des.* **19,** 2499–2511 (2013).
138. Hopkins, S. C. *et al.* Pharmacodynamic effects of a D-amino acid oxidase inhibitor indicate a spinal site of action in rat models of neuropathic pain. *J. Pharmacol. Exp. Ther.* **345,** 502–511 (2013).
139. Terry-Lorenzo, R. T. *et al.* Novel human D-amino acid oxidase inhibitors stabilize an active-site lid-open conformation. *Biosci. Rep.* **34,** e00133 (2014).

140. Todone, F. *et al.* Active site plasticity in D-amino acid oxidase: a crystallographic analysis. *Biochemistry* **36,** 5853–5860 (1997).
141. Somerville, A. N. SciFinder Scholar. *J. Chem. Educ.* **75,** 959, 975–976 (1998).
142. Sparey, T. *et al.* The discovery of fused pyrrole carboxylic acids as novel, potent D-amino acid oxidase (DAO) inhibitors. *Bioorg. Med. Chem. Lett.* **18,** 3386–3391 (2008).
143. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78,** 1950–1958 (2010).
144. Dupradeau, F.-Y. *et al.* RE DD. B.: a database for RESP and ESP atomic charges, and force field libraries. *Nucleic Acids Res.* **36,** D360–D367 (2007).
145. Duplantier, A. J. *et al.* Discovery, SAR, and pharmacokinetics of a novel 3-hydroxyquinolin-2 (1 H)-one series of potent d-amino acid oxidase (DAAO) inhibitors. *J. Med. Chem.* **52,** 3576–3585 (2009).
146. Wichapong, K., Nueangaudom, A., Pianwanit, S., Tanaka, F. & Kokpol, S. Molecular dynamics simulation, binding free energy calculation and molecular docking of human D-amino acid oxidase (DAAO) with its inhibitors. *Mol. Simul.* **40,** 1167–1189 (2014).
147. Katane, M. *et al.* Identification of novel D-amino acid oxidase inhibitors by in silico screening and their functional characterization in vitro. *J. Med. Chem.* **56,** 1894–1907 (2013).
148. Kawazoe, T., Tsuge, H., Pilone, M. S. & Fukui, K. Crystal structure of human D-amino acid oxidase: Context-dependent variability of the backbone conformation of the VAAGL hydrophobic stretch located at the si-face of the flavin ring. *Protein Sci.* **15,** 2708–2717 (2006).
149. Adage, T. *et al.* In vitro and in vivo pharmacological profile of AS057278, a selective d-amino acid oxidase inhibitor with potential anti-psychotic properties. *Eur. Neuropsychopharmacol.* **18,** 200–214 (2008).