SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Earth and Environmental Sciences

PhD program Chemical, Geological and Environmental Sciences Cycle XXXIV

Curriculum in Chemical Sciences

MODELING OF LIGAND-PROTEIN BINDING WITH ADVANCED MOLECULAR DYNAMICS METHODS

Callea Lara

Registration number 748402

Tutor: Prof. Claudio Greco

Supervisor: Prof.ssa Laura Bonati

Coordinator: Prof. Marco Malusà

ACADEMIC YEAR 2020/2021

To my sweet dad,

to you, my dear Ciccillino.

“Something that is loved is never lost.”

Toni Morrison

CONTENTS

ABBREVIATIONS

Protein

aDHS	archeal Deoxyhypusine Synthase
AhR	Aryl hydrocarbon Receptor
ARNT	Aryl hydrocarbon Receptor Nuclear Translocator
DHS	Deoxyhypusine Synthase
eIF5A	eukariotic Initiation Factor 5A
hDHS	human Deoxyhypusine Synthase
HIF-2α	Hypoxia Inducible Factor 2α
MAPK	Mitogen-Activated Protein Kinase
VEGF	Vascular Endothelial Growth Factor

Computational Techiques

BOMB	Born-Oppenheimer molecular dynamics
CG	Coarse Grained
CG-MD	Coarse-Grained Molecular Dynamics
CM	Continuum Mechanics
CP-MD	Car-Parrinello Molecular Dynamics
DFT	Density Functional Theory
FEP	Free Energy Perturbation
GAFF	Generalized Amber Force Field
GaMD	Gaussian accelerated Molecular Dynamics
Glide XP	Glide Extra Precision
HF	Hartree-Fock
InMetaD	Infrequent Metadynamics
KS-DFT	Kohn-Sham Density Functional Theory
LIE	Linear Interaction Energy
MD	Molecular Dynamics
MetaD	Metadynamics
MM	Molecular Mechanics
MM-GBSA	Molecular Mechanics Generalized Born Surface Area
MM-PBSA	Molecular Mechanics Poisson-Boltzmann Surface Area
MR	Molecular Recognition

MSMs	Markov State Models
PathDetect-SOM	Pathways Detection on SOM
PBC	Periodic Boundary Condition
PCVs	Path Collective Variables
PME	Particle Mesh Ewald
PP	Physical Pathway
QM	Quantum Mechanics
QM/MM	Quantum Mechanics/Molecular Mechanics
RESP	Restricted Electrostatic Potential
sMD	steered Molecular Dynamics
SOM	Self-Organizing Map
TI	Thermodynamic Integration
US	Umbrella Sampling
WT-MetaD	Well-Tempered Metadynamics
τ-RAMD	τ-Random Acceleration Molecular Dynamics

General

bHLH	basic Helix–Loop–Helix
BMU	Best Matching Unit
Cryo-EM	Cryogenic Electron Microscopy
CV	Collective Variable
dRMSD	distance Root Mean Square Deviation
FES	Free Energy Surface
HPC	High Performance Computing
MiMiC	Multiscale Modeling in Computational Chemistry
NMR	Nuclear Magnetic Resonance
PAS	PER-ARNT-SIM
PDB	Protein Data Bank
PES	Potential Energy Surface
PMF	Potential of Mean Force
RMSD	Root Mean Square Deviation
RMSF	RMSF Root Mean Square Fluctuation
VMD	Visual Molecular Dynamics

INTRODUCTION

Many of the biological mechanisms that regulate the life of living organisms involve "*molecular recognition (MR)*"[1,2] processes in which small molecules (ligands) bind a specific region (binding site) of targeted macromolecules, such as proteins, through non-covalent interactions to form a complex. Knowledge of ligand-protein interaction mechanisms is an essential prerequisite in the design, discovery, and development of new drugs, by helping to determine when a molecule has the potential to become a drug[3,4]. Two different aspects need to be taken into account when studying the ligand binding process and designing new drugs: the thermodynamic one, which involves identification of the correct binding mode and estimation of the binding free energy (associated to binding-affinity), and the kinetic one, *i.e.* determination of the kinetic constants for binding and unbinding as well as of the activation barriers and the rate determining step. A new drug should be designed to improve interactions with the binding site and to have, at the same time, a good kinetic binding profile[5].

1.1 Thermodynamics and kinetics of ligand-protein binding

The ligand-protein binding event is governed by the rules of a simple reversible reaction[5], in which a protein (P) and a ligand (L) in their free form in solution bind and form the PL complex in solution.

When the equilibrium state is reached the association (K_a and the dissociation (K_d) constants can be calculated from the ratio:

$$K_a = \frac{[PL]}{[P][L]} = \frac{1}{K_d}$$

where [PL], [P] and [L] are the equilibrium concentrations of the three components. At equilibrium, the rates of the forward (binding) and the reverse (unbinding) reactions are balanced, and the association constant (K_a) can be expressed as the ratio of the kinetic constants:

$$K_a = \frac{k_{on}}{k_{off}}$$

Likewise, from a thermodynamic point of view, at constant temperature and pressure (as in biological systems), a spontaneous process (such as ligand binding) takes place only if a negative change in the Gibbs free energy occurs. Under standard conditions, the relationship between K_a and the difference in Gibbs free energy of the system between the unbound and bound state (ΔG_{bind}, binding free energy), is given by:

$$\Delta G_{bind} = -RTln(K_a)$$

where R is the universal gas constant and T is the temperature. The complex will be more stable when the ΔG_{bind} is more negative.

The binding free energy can also be expressed as a function of the changes in enthalpy (ΔH) and entropy (ΔS) on formation of the complex:

$$\Delta G = \Delta H - T\Delta S$$

Changes in enthalpy arise from the breaking or the formation of non-covalent interactions between the protein, the ligand, and the solvent molecules, and from

conformational changes of protein and ligand upon complexation. On the other hand, changes in entropy reflects the change in translational and rotational degrees of freedom for the ligand, protein and solvent molecules, and the loss of conformational entropy of P and L on binding. From a thermodynamic point of view, the free-energy is a state function, implying that the variation of free-energy in the process depends only on the free-energy of the two states considered and not on the path taken to connect them. This implies that the energy of the bound and unbound states are sufficient for the calculation of binding thermodynamic quantities. The kinetics of a process, on the other side, is related to the height of the energy barrier to overcome according to the Eyring equation:

$$k = \frac{k_b T}{h} e^{-\frac{\Delta G^{\ddagger}}{RT}}$$

Where k is the velocity constant, $\Delta G^{\ddagger}$ is the free energy of activation, i.e. the difference between the free energy of the transitions state and that of the reacting systems, k_b is the Boltzmann's constant, h is the Planck's constant and T the absolute temperature. Higher is the energy barrier to overcome, less frequently a transition is observed. The study of ligand binding kinetics has become of particular interest, since often drug efficacy correlates better with binding kinetics than with the binding affinity alone[6].

1.2 Molecular modelling for studying ligand-protein interactions

Over the years, many computational methods[7] have been developed to address the study of ligand-protein binding, ranging from simple molecular docking procedures to computationally demanding but more accurate methods based on both classical- and quantum-mechanics. Thanks to the development of new computer technologies and of new computational techniques, *in silico* methods are becoming increasingly effective for calculating the thermodynamic and kinetic properties underlying ligand-protein interaction mechanisms by providing insights of the process at the atomistic level.

Computational methods based on molecular mechanics (MM)

Ligand-binding can be described by using methods based on Molecular Mechanics (MM). Among these, Molecular Docking estimates the binding free energy of the binding poses by means of scoring functions mostly derived from MM potentials. The first model to describe ligand-protein binding was the *lock and key* model proposed by Fisher and based on the perfect matching of the binding surfaces of the ligand and the protein, which are considered to be rigid bodies. This model was adopted in the first molecular docking approaches,[8] in which both the ligand and the protein structures are kept rigid during sampling of the conformational space of the system to predict the correct binding mode.[9]. The power of using such simple docking techniques is the speed of the calculation that allows the screening of millions of compounds with an affordable computational cost[10]. However, they provide a static representation of the system and are neither able to provide mechanistic information nor to explain the transmission of the effects caused by binding. Proteins are dynamic objects and the protein flexibility associated with ligand binding plays a crucial role in the correct prediction of the binding mechanism and the related kinetic and thermodynamic properties. Different models were proposed to explain the role of protein conformational dynamics in these processes. The *induced fit* model relies on the hypothesis that binding of a ligand to the protein binding site induces a change in the shape of the protein for a reciprocal structural adaptation. In the *conformational selection* model, the ligand selects the most complementary protein conformation from an ensemble of pre-existing metastable states, which in turn shifts the dynamic population equilibrium toward the conformation adopted in the bound state [11]. The above models have been used by a number of computational strategies. Some of them, related to the induced-fit model, include the effects of conformational variation of the protein in docking calculations: soft-docking, in which the repulsion terms between the binding site of the protein and the ligand are attenuated; treatment of the conformational freedom of sidechains at the binding site during sampling; refinement of the docking poses by Molecular Dynamics (MD) simulations. A method that relies on the conformational selection model is ensemble docking, in which docking is not

performed to a single protein conformation but to a set of conformations that can be derived either from experimental structures (obtained with X-ray or NMR) or from MD simulations[12,13].

Looking over the use of MD to include flexibility in docking calculations, nowadays it is possible to perform MD simulations to explore the free energy landscape and the kinetic profile associated with the investigated process, thus obtaining a complete dynamic description of the ligand-protein binding event. With the current computational power, it is possible to produce atomistic simulations up to milliseconds on specialized architectures such as the Anton supercomputer, or tens/hundreds of microseconds on standard high performance computing (HPC) facilities. This means that nowadays it is even possible to observe rare events (such as ligand binding) that happen on the micro/millisecond timescales using MD. Helped by the exponential increase in the computational power, in recent years several methods based on MD simulations have become increasingly popular for describing the ligand-protein binding. These methods can be classified in two categories: those aimed at estimating the ligand-protein binding affinity by providing information about thermodynamic properties; and those focused on a clearer understanding of the complete process by providing information about the ligand binding (and/or unbinding) pathways[14]. The first category includes *end-state methods*, based on the property of free energy to be a state function, and thus focused on characterization of the bound and unbound states. These methods are usually used in a post-processing manner, where the free energy is estimated based on MD simulations. Examples are: Linear Interaction Energy (LIE)[15]; Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA)[16] and Molecular Mechanics Generalized Born Surface Area (MM-GBSA)[17]; alchemical free-energy perturbation methods, such as Thermodynamic Integration (TI)[18] and Free Energy Perturbation (FEP)[19]. Methods that fall in the second category are defined *physical pathway* (PP) *methods* and enable the simulation of the complete binding and/or unbinding events, which can lead to calculation of both thermodynamic and kinetic properties, and to characterization of energy barriers and relevant intermediate states along the pathways. As mentioned above, with MD it is now possible to observe ligand

binding event. However, computation of key thermodynamic quantities requires the observation of multiple binding events to obtain reliable statistics on the process, thus increasing the computation time. To this aim, *enhanced sampling methods* need to be employed; [7,20] they are based on innovative algorithms to overcome the sampling issue by speeding up the description of slow processes. Examples of these methods are Steered Molecular Dynamics (sMD)[21,22], Umbrella Sampling (US)[23,24], Metadynamics (MetaD)[25] and its different variations[25–28], Gaussian accelerated MD (GaMD)[29], scaled MD[30,31], τ-RAMD[32], MD Binding[33,34] and CG-MD[35].

The methods discussed up to now, from molecular docking to MD with enhanced sampling techniques, are based on molecular mechanics, whereby electrons are not explicitly included in the calculations, and the energy of molecules is calculated using parametrized classical potentials (force fields, FFs). The advantage is that this speeds up the calculations considerably, allowing extended MD simulations of large biomolecular systems. On the other hand, there are several limitations to the use of classical FFs in MD simulations. For example, in most of the force-field, the continuous electron distribution is usually approximated with fixed point charges centered on the atomic nuclei. Only few FFs include a description based on dipoles or multipoles[36] and/or allow the atomic charge to be modified as a result of polarization or charge-transfer effects occurring during the process[37,38]. Another limitation is the impossibility of correctly studying chemical reactivity since chemical bonds break or formation cannot be described in a classical MD simulation[7]. To address these issues, methods based on quantum mechanics (QM) can be used.

Computational methods based on quantum mechanics (QM)

QM methods are based on first principles and explicitly consider both nuclei and electrons. For these reasons, they are more accurate for studying properties, interactions and processes in biomolecular systems[39]. These methods aim to solve the electronic Schrödinger equation given fixed positions of the nuclei in order to obtain useful information about a molecular system, such as electron densities and related electronic properties, equilibrium geometries and energies. Moreover, they allow a

realistic description of the intermolecular interactions and reactivity of the systems[40]. To achieve this, different methods which differ in computational cost and accuracy can be used, such as Hartree-Fock (HF), Density Functional Theory (DFT) and semi-empirical (SE) methods[41,42]. Nowadays, QM approaches can be used also to perform molecular simulations, allowing the treatment of systems with appreciable size and complexity over ps time scale, but they require much more computational resources than MM approaches. This is why QM models are mostly applied to static structures and the inclusion of sampling of dynamic processes is still an open and ongoing challenge[39,43]. On the other side, as outlined above, classical MD simulations allow very long MD trajectories to be calculated for millions of atoms, but the quality of the simulation depends on the accuracy of the empirical function used and many important properties such as polarization, charge transfer and bond breaking/formation cannot be described. A valid strategy to perform sampling of the ligand-binding process and to achieve a good level of accuracy with acceptable computational costs is to use multiscale hybrid approaches combining different levels of theory[44,45]. In particular, these methods take advantage of the mixed ability of the selected methods to treat a specific region: usually the ligand or the binding site are treated at a high level of theory (QM) in order to get a more accurate description of the most interesting part of the system, while the remaining part is treated at a lower level (MM). Currently, thanks to the improvement of the computational power, hybrid QM/MM MD simulations are being employed to study biomolecular systems[46,47] and the application of QM/MM approaches for drug design is a wide and rapidly growing field.

1.3 Motivation and thesis outline

From what has been discussed above, it emerges that computational methods based on molecular dynamics are becoming attracting for the description of the ligand binding event. However, some major problems have to be taken into consideration.

Physical pathway methods allow the simulation of the complete binding and/or unbinding process and may lead to calculation of the related thermodynamic and kinetic properties. However, ligand binding events occur in time scales inaccessible by conventional molecular dynamics and therefore computational approaches based on enhanced-sampling methods have to be used to speed up the simulation. The large amount of data generated by extensive sampling requires appropriate tools to analyze the simulated events and to provide a clearly interpretable picture.

While simulations based on classical MM potentials suffer from some limitations, the introduction of a higher level of theory such as QM allows to describe some important features of molecular interactions. Hybrid QM/MM methods allow to combine the advantages of both methods, namely to obtain a more accurate description of the process and to retain low computational costs.

In this PhD thesis, several issues related to the above topics are addressed.

In Chapter 2 (Theoretical basis of Molecular Dynamics Simulations), a theoretical introduction to molecular dynamics simulations at both the MM and QM/MM levels is given. In addition, a more extensive description of the methods used in the PhD project for both levels of theory is given. The computational details related to the approaches used for the different studies are presented and discussed in the corresponding chapters.

In Chapter 3 (Metadynamics-Based Approaches for Modeling the Hypoxia-Inducible Factor 2α Ligand Binding Process), advantages and limitations of using several enhanced sampling methods for the description of ligand-binding are analyzed. An approach based on the combination of the more efficient methods is proposed to investigate and predict the possible binding/unbinding pathways of the ligand and to obtain a correct estimation of the binding free energy.

In Chapter 4 (PathDetect-SOM: A Neural Network Approach for the Identification of Pathways in Ligand Binding Simulations), the difficulty of analyzing large amounts of data from several replicas or from a single simulation describing several re-crossing events obtained by enhanced sampling methods is addressed. Hence a tool based on

Neural Networks is proposed to analyze all simulated events at the same time and to provide a clearly interpretable overview of the differences in the sampled pathways.

In Chapter 5 (Investigation of ligand-protein interaction through highly scalable QM/MM MD simulations), the advantages of introducing QM approaches for the description of the ligand and its interactions are analyzed. In particular, the improvement in the description of key properties such as polarization of the electron density using mixed QM/MM approaches via the MiMiC interface is explored.

Finally, in Chapter 6 (Conclusions), some conclusive remarks related to the application of different computational methods for the study of ligand-protein binding are reported, and the main results deduced from the complete PhD project are summarized.

THEORETICAL BASIS OF MOLECULAR DYNAMICS SIMULATIONS

MD simulations provide an opportunity to study the dynamics behaviour of proteins at an atomistic level, thereby helping to understand biological mechanisms. A description of the chemical and physical ligand-protein interactions can be obtained using different approaches, on the basis of the dimension of the biomolecular system to be investigated and of the accuracy required: classical MD, hybrid QM/MM MD or QM MD.

This chapter presents the theoretical basis of the main computational methods used in the thesis. In particular, the first part focuses on: the fundamentals of the molecular mechanics (MM) approximation; the theory of classical MD simulations; and some enhanced sampling methods used in this PhD project, such as steered molecular dynamics (sMD) and metadynamics (MetaD) with the Path Collective Variables (PCVs) formalism. The second part provides the theoretical background to carry out a QM/MM study of a ligand-protein system and to perform MD simulations at this level of theory.

2.1 Classical approach for MD simulations

Molecular Mechanics (MM)

Molecular mechanics (MM)[48] is based on a "sphere and spring" model of a molecule, in which atoms are treated as spheres connected by springs that represent the bonds. The electronic structure of the molecule is neglected, and the energy of the molecule is approximated by the terms of a force field. The force field is an empirical potential energy function that describes the energy of a molecule as a function of the Cartesian coordinates of all atoms. A standard force field for biomolecular simulations usually consists of a sum of different terms:

$$U = U_{bonds} + U_{angles} + U_{torsions} + U_{el} + U_{vdW}$$

The first three terms describe the internal energy of the molecule, coming from all bonds, angles, and dihedrals present in the molecule, whereas the last two terms are the non-bonded terms. In most variants of MM, covalent bonds are represented by springs, so the first term, U_{bonds}, uses the harmonic potential to describe covalent bond stretching around the equilibrium bond length, r_{eq}:

$$U_{bonds} = \sum k_b(r - r_{eq})^2$$

where k_b is the spring force constant for a given bond type b. Similarly, the bond angle term, U_{angles}, is also described by a harmonic potential:

$$U_{angles} = \sum k_a (\theta - \theta_{eq})^2$$

where k_a is the angle force constant for angle a involving three bonded atoms, and θ_{eq} is the equilibrium angle. The dihedral term, $U_{torsions}$, usually describes the torsion angle rotation around a bond with a periodic function:

$$U_{torsions} = \sum \frac{V_n}{2}[1 + \cos(n\phi - \phi_{eq})]$$

where, V_n is the corresponding force constant, n is the periodicity of the rotation, and ϕ_{eq} is the equilibrium angle.

The non-bonded terms are usually computed between atoms separated by more than three bonds, to exclude from the computation the pairs of atoms included in the internal terms.

The Coulomb electrostatic interaction energy between two atoms *i* and *j* with partial atomic charges q_i and q_j is computed as:

$$U_{el} = f\frac{q_i q_j}{\varepsilon_r r_{ij}}$$

where f is the electric conversion factor $\left(\frac{1}{4\pi\varepsilon_0}\right)$, ε_r is the dielectric constant of the given medium, and r_{ij} is the interatomic distance between atoms *i* and *j*.

The other non-bonded term is the van der Waals (vdW) potential, that describes the dipole–dipole interactions, including the dispersive interactions between instantaneous dipoles (London forces). At large interatomic distances, this term should be equal to zero, whereas at very short distances it should be strongly repulsive. However, at intermediate distances, where atoms are close to each other, but their electron clouds are not overlapping, this term should be slightly negative. This behaviour is well described by the Lennard-Jones (LJ) potential, which consists of two parts, a short-range repulsive term, and a long-range attractive term:

$$U_{vdW} = \sum_{i<j} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$

where ε_{ij} corresponds to the depth of the potential energy curve, σ_{ij} is the finite distance between two atoms at which U_{vdW} is zero, and r_{ij} is the distance between the two atoms (Figure 2.1).

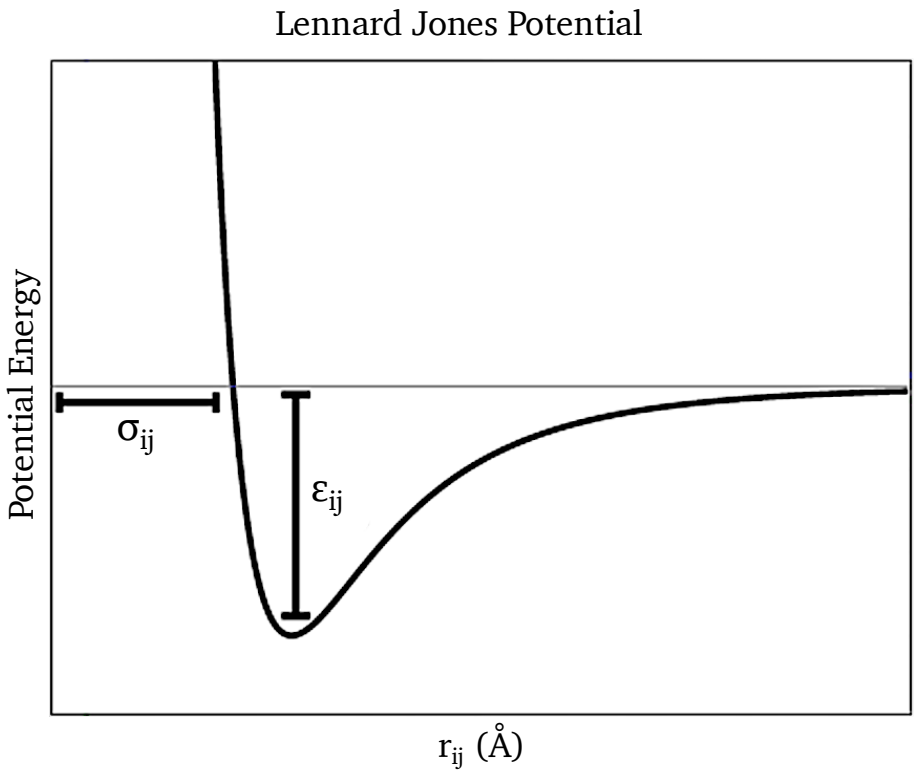

Figure 2.1 Shape of the Lennard Jones potential function.

Classical molecular dynamics (MD)

Molecular dynamics (MD) is a computer simulation technique that describes the evolution of a system over the time using the force field equation, discussed in the previous section. This method plays an important role in drug design and discovery because it allows processes involving very complex systems to be studied at the atomic level starting from a static structure produced by X-ray crystallography, nuclear magnetic resonance (NMR), Cryogenic electron microscopy (Cryo-EM) or homology modeling. Indeed, it is known that considering the flexibility of the protein, and thus its constant movement, rather than a single frozen structure allows a better understanding of the ligand-protein binding process[49,50].

In MD simulations, the trajectory is obtained by solving the differential equation derived from the Newton's second law:

$$\frac{d^2 x_i}{dt^2} = \frac{F_{x_i}}{m_i}$$

where F_{x_i} is the force acting on the particle i with mass m_i moving along one coordinate x_i. MD simulations start by assigning initial positions $r(t)$ and velocities $v(t)$ (the first derivative of the positions with respect to time) to all particles in the system. By integrating the Newton's equations of motion, it is possible to calculate the corresponding positions and velocities at time $t + \Delta t$, where Δt is the time step used in the simulations. For a small time step, forces may be considered constant and it is possible to solve the equations of motion, where the positions of the particles in the system can be approximated by a Taylor series expansion:

$$r(t + \Delta t) = r(t) + v(t)(\Delta t) + \frac{1}{2}a(t)(\Delta t)^2 + \frac{1}{6}b(t)(\Delta t)^3 + \cdots$$

$$v(t + \Delta t) = v(t) + a(t)(\Delta t) + \frac{1}{2}b(t)(\Delta t)^2 + \frac{1}{6}c(t)(\Delta t)^3 + \cdots$$

$$a(t + \Delta t) = a(t) + b(t)(\Delta t) + \frac{1}{2}c(t)(\Delta t)^2 + \cdots$$

where r is the position, v is the velocity, a is acceleration (the second derivative of the positions with respect to time) obtained from the calculated force acting on each atom, and b and c respectively the third and fourth derivative. After calculating the new positions of all atoms, it is possible to update the energy and forces and iterate the procedure as many times as needed. This way, an MD simulation produces a trajectory, which shows how positions and velocities of all atoms vary with time. As discussed before, the forces may only be considered constant if the time step used is small. Special attention must be paid to the time step used: if it is too small, the trajectory will progress slowly, thus covering only a limited region of the phase space, while if it is too large high energy atoms may overlap causing instability in the integration algorithm. A useful rule for establishing the time step to be used is that it should be one-tenth the time of the shortest period of motion. In MD simulations of biomolecular systems, the highest frequency motion is the stretching of bonds involving hydrogen atoms which vibrates with a period of about 10 fs. Accordingly, the recommended time step is 1 fs. A possible strategy to allow the use of larger time step is to constrain bonds involving hydrogen atoms to their equilibrium values using the SHAKE[51] or LINCS[52] algorithms.

Given that the size of the system to simulate cannot be infinite, usually the molecules of interest are embedded in a box of finite size. In order to be able to calculate macroscopic properties from a MD simulation, it is necessary to treat the boundaries and boundary effects correctly. This becomes particularly important when simulations are performed by explicitly treating solvent molecules. Considering a finite system during simulation, it is necessary to use walls that limit the diffusion of molecules; artefacts can be observed in proximity to these walls. A method to overcome this problem is the use of the periodic boundary condition (PBC) approach (Figure 2.2).

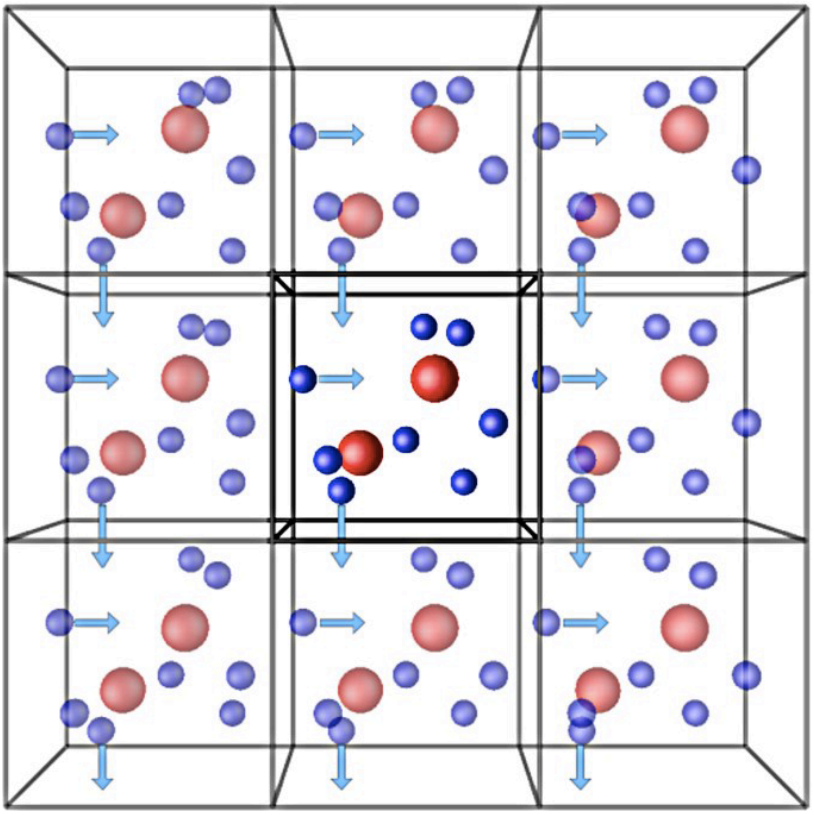

Figure 2.2 Schematic representation of periodic boundary conditions.

With this approach, a unit box of the system is replicated in all directions forming periodic images of the same box; the particles in the adjacent boxes will move in the same way as those in the original box. During the simulation, all boxes are identical and when a particle leaves the original box, a particle enters the box from the opposite side, ensuring that the total number of particles in the original box is preserved. The unit cell must have a shape that can be replicated in 3D space by forming a lattice

without holes. If periodic boundary conditions are used, interactions between the system in the central box and its periodic images should also be taken into account.

The number of non-bonded interactions exhibit a quadratic growth with the number of atoms. This mean that for very large systems, the number of interactions to be computed become prohibitive. VdW interactions decay very quickly and therefore can be cut to a certain cutoff (typically 9-12 Å) reducing the number of interactions to be considered. However, for electrostatic interactions ignoring interaction between atoms beyond the cutoff value is not appropriate as it introduces serious errors in the force calculation. Indeed, given that the energy of these interactions at the cut-off distance are not completely negligible, methods to overcome this issue should be used such as the *Particle Mesh Ewald* (PME)[53] approach. The PME method uses a summation in the Fourier space for the long-range part which quickly converges in the Fourier reciprocal space.

2.2 Enhanced Sampling Methods

Enhanced sampling methods aim to improve the sampling of classical MD simulations for biomolecular system. Some events involving this type of systems, such as ligand binding, occur in time scales not accessible with classical MD simulations. In fact, the integration time step is limited to femtoseconds and therefore, to observe events in micro/milliseconds timescale, computations of billions or trillions of steps are required. An important bottleneck for an efficient exploration of the phase space during MD simulations is the presence of high energy barriers separating different conformations leading to the rare observation of transitions between them. Therefore, simulations of complex systems require high computational costs. For this reason, enhanced sampling methods are required to accelerate the sampling of the conformational space. Using these approaches, it is possible to calculate thermodynamic and/or kinetic properties associated with the process of interest. The central idea of most of the enhanced sampling methods is to add a bias potential to the Hamiltonian of the system, to allow it to overcome the free energy barrier and

consequently to sample new regions. Many of these methods use predefined reaction coordinates or collective variables (CVs) to guide simulations effectively[54,20,55,56].

In the following paragraphs the enhanced sampling methods used in this work are discussed in detail, in particular steered MD and Metadynamics with the Path Collective Variables formalism.

Steered molecular dynamics (sMD)

Steered molecular dynamics (sMD)[21,57,58,59] is a non-equilibrium method in which a time-dependent external force is applied to a specific set of atoms and pulls the system along the reaction coordinate (or collective variable, CV) to facilitate the sampling (Figure 2.3).

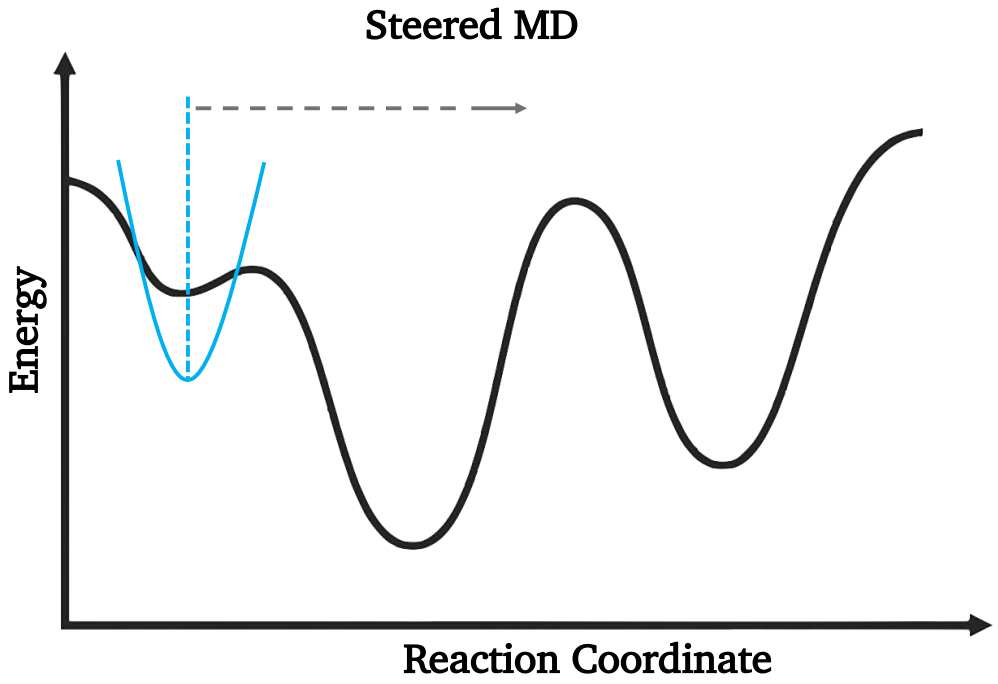

Figure 2.3 Schematic explanation of steered MD in which the bias is moved along a specific reaction coordinate.

In other terms, sMD acts by pulling the system along one or more CVs to guide the system from an initial configuration to a final one (for ligand binding from the bound state to the unbound state). In particular, at the beginning of the simulation a harmonic time dependent potential U acting on a selected CV, is added to the standard Hamiltonian:

$$U = \frac{K}{2}(x - x_0)^2$$

where K is the strength of the external force applied (spring constant) and x_0 is the initial value of the CV. The force F applied to the system during the simulation can be expressed as:

$$F = K(x_0 + vt - x)$$

where v is the pulling velocity. Finally, the external work ΔW performed on the system is derived from the F by integrating the force over the pulled trajectory:

$$\Delta W = v \int_{t_0}^{t_f} F(t)\, dt$$

ΔW represents the cumulative change of the Hamiltonian in time and it is related to the change in energy of the system. With the Jarkzynski equality it is possible to connect the free energy difference between two states with the ensemble average of work obtained with sMD simulations:

$$\Delta G = -\frac{1}{\beta} ln\langle e^{-\beta W}\rangle$$

where $\beta = \frac{1}{k_B T}$, with T and k_B being the temperature of the system and the Boltzmann constant, respectively.

In summary, in sMD simulations the equilibrium free energy of the system is obtained from the average of the irreversible works. Consequently, it is necessary to perform a large number of independent replicas to provide a statistically significant calculation of W and in turn a reliable estimate of the free energy. However, given that Jarkzynski equality involves the average of a noisy quantity (the work) that appears in the exponential, calculation of $\langle e^{-\beta W}\rangle$ leads to large errors. It is possible to observe that a higher irreversible work is associated with a higher variance, consequently, the optimization of the force applied to the ligand along its unbinding, or the application of different pulling velocities, leads to an improvement of the convergence of the results[14].

Metadynamics (MetaD)

Metadynamics (MetaD)[60,26,61], is a well-known enhanced sampling method for studying rare events. In MetaD simulations a history-dependent bias potential is added in order to enhance the sampling by depositing a bias with a Gaussian shape distribution, at regular time intervals on the current position of the selected CVs (Figure 2.4).

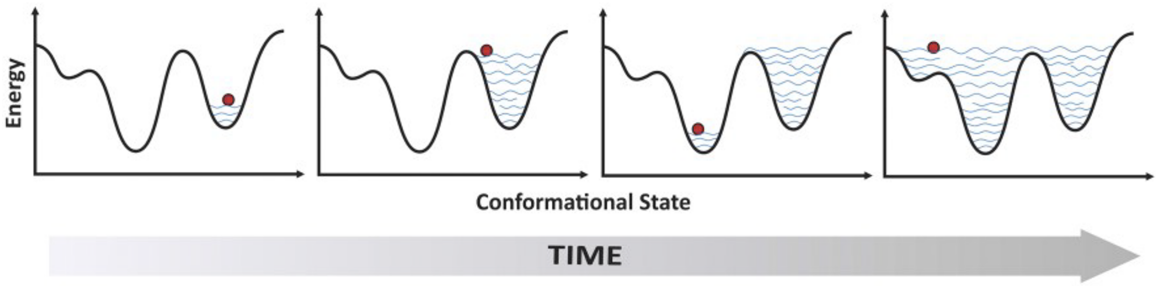

Figure 2.4 Schematic explanation of metadynamics in which an history-dependent bias is added to discourage re-sampling. (Image modified from ref [20])

The bias potential acting on the system at time t is expressed by means of the following function:

$$V_G(S(x),t) = w \sum_{\substack{t'=\tau_G,2\tau_G,\dots \\ t'<t}} e^{\left(-\frac{\left(S(x)-s(t')\right)^2}{2\delta s^2}\right)}$$

where w is the Gaussian height, τ_G is the frequency of the Gaussian deposition along the CV, δ_s is the Gaussian width and $s(t) = S\big(x(t)\big)$ is the value taken by the CV at time t. These parameters determine the accuracy and efficiency of the free energy reconstruction. If the Gaussians are large, the free energy surface (FES) will be explored at a fast pace, but the reconstructed profile will be affected by large errors. Instead, if the Gaussian are small or are placed infrequently the reconstruction will be more accurate, but it will take a longer time. During the simulation, the bias potential fills the the FES minima, allowing the system to efficiently explore the space defined by the CVs. The choice of CVs is very important since the efficiency of the method scales exponentially with the number of dimensions involved. Ideally the CVs should satisfy three properties[60]: the first is that they should clearly distinguish between the initial state, the final state, and the intermediates state; the second is that they should

describe all the slow motion relevant for the process and the last one is that their number should not be too large, otherwise it will take a very long time to fill the FES.

The main drawback of using MetaD is related to assessing the convergence of the simulation. Once all the basins are visited, and while the simulation keeps running, the bias continues being deposited. This has the effect of overfilling a minimum and the height of the accumulated Gaussians will largely exceed the true barrier height (hysteresis). Thus, for a reliable FES estimate, the simulation should be stopped as soon as the system starts diffusing in the CVs space. A solution to this problem is provided by well-tempered metadynamics (WT-MetaD)[62]. While in standard MetaD Gaussians of constant height are deposited over time, in WT-MetaD the Gaussian height becomes a function of the simulation time and it is scaled by a factor:

$$e^{-\frac{V(s(t),t)}{k_B \Delta T}}$$

where the bias potential has been evaluated at the same point where the Gaussian is centered and ΔT is the range of the temperature at which the CVs are sampled. When the system reaches a new basin, the initial Gaussian height is reset, and the scaling of the hills starts again. As a consequence, the bias potential tends to converge smoothly in the long time limit. The choice of the entity of Gaussian height decrease per unit time is very important. However, the Gaussian height should not become too small before a basin is completely filled, otherwise the system would be trapped inside the basin with no possibility of overcoming the barriers. This can be monitored by setting a specific parameter for the simulation, the *biasfactor*, defined as:

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

where ΔT is the upper limit of the temperature range to which the sampling of the CVs is confined. Thus, in the long-time limit, the system will explore the biased canonical distribution:

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

where $F(s)$ is the free energy. Due to the bias potential, CVs explore the canonical ensemble at an effective temperature $T+\Delta T$. WT-MetaD is therefore of great advantage because it allows the exploration of CV space to be confined only to regions of reasonable free energy.

As previously explained, another important issue encountered when using MetaD is the identification of an appropriate set of CVs, else the simulation will not converge and the system will remain stuck in a certain position until the rare event involving the hidden CV eventually occurs. 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 path collective variables (PCVs).

Path Collective Variables (PCVs)

The idea of approximating the reaction coordinate with a path connecting two stable basins in energy or free energy space is useful in clarifying reaction mechanisms. Path variables can be used in metadynamics to effectively overcome the difficulty of managing highly dimensional phase spaces and reduce the choice between multiple possible CVs.

Path Collective Variables (PCVs) formalism, proposed by Branduardi et al[28], involves a set of path variables that have been successfully used to investigate complex chemical and biological processes and compute their associated free energy surfaces and kinetics[63,64,65,66,67]. The idea behind this approach is the possibility of describing a transition between state A and B with a series of frames (frameset) capturing the system at intermediate states along the reaction coordinate (Figure 2.5).

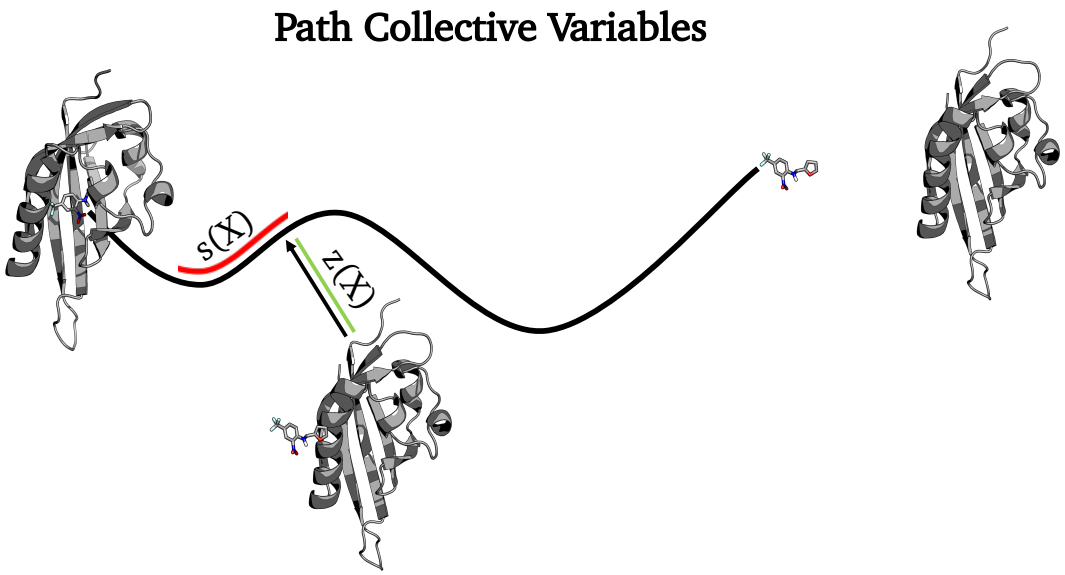

Figure 2.5 Schematic explanation of path collective variables in which sampling along the path is guided by s(X) and z(X).

This frameset can be used to guide the sampling along the path by two different CVs: the first one, s(X) describes the progression along the path, while the second one, *z(X)*, is orthogonal to s*(X)* and expresses the distance from the path. While s*(X)* describes the evolution along the pathway, z*(X)* allows to explore adjacent regions of the phase space. The formalism of the two CVs can be expressed as:

$$s(X) = \frac{\sum_{i=1}^{N} i\, e^{(-\lambda R[X-X_i])}}{\sum_{i=1}^{N} e^{(-\lambda R[X-X_i])}}$$

$$z(X) = -\frac{1}{\lambda} ln\left(\sum_{i=1}^{N} e^{(-\lambda R[X-X_i])}\right)$$

where X represent the atomic coordinates at the current simulation time step, while X_i denotes those of the i-th frame with N the total number of frames comprised in the frameset. The difference $X - X_i$ is the distance between configuration X and the one adopted in frame i. The function R is the chosen metric which measures this distance. The λ parameter serves to smooth the variation of the variables and obtain a continuous collective variable. A good starting value for the tuning of its value can be obtained with the following formula:

$$\lambda = \frac{2.3(N-1)}{\sum_{i=1}^{N-1} |X_i - x_{i+1}|}$$

which imply that the average distance between consecutive frames composing the path is calculated. As mentioned above, different metrics can be used for calculating distances. In the original implementation of PCVs, the Root Mean Square Deviation (RMSD) was the chosen metric, requiring the alignment of structures to the reference path at each time step. An interesting alternative is to use the distance-RMSD, or dRMSD, which measures the differences between atomic distances within structures. The dRMSD metric avoids the problem of structure alignment. In summary, the highly dimensional space is reduced to a description exploiting the progression along the pathway as CV. The main advantage from combining $s(X)$ to $z(X)$ is a more complete exploration of the conformational space. There is no general rule to obtain the required frameset which provides the reference path, but it is necessary to consider some crucial aspects: first of all, consecutive frames need to describe unidirectional progression towards the final state; secondly, equal spacing between frames is required; and finally, an appropriate number of frames should be chosen, so that the distance between subsequent frames should not be excessive[68]. Finally, it should be considered that it is difficult to distinguish conformations that are similarly "distant" from conformations that are dissimilar to the reference structure. A good path, thus, should provide a free-energy surface with the deepest minima not too distant from the reference path in $z(X)$; going too far from the reference path increases the probability of finding different states with the same CVs values.

2.2 QM/MM approach for MD simulations

Molecular mechanics methods have the advantage of being computationally accessible but also have some limitations as they completely ignore the electronic structure and describe the molecule by treating atoms as spheres and bonds between them as springs. In contrast, quantum mechanical methods explicitly include electronic treatment, thus enabling the processes under investigation to be studied more accurately; however, the major limitation is that they are computationally onerous, and it is therefore difficult to study large biomolecular systems, such as ligand-protein complexes, at this level of theory. A solution to this problem are the hybrid QM/MM

methods. The original QM/MM method was implemented by Warshel and Levitt to explore the catalytic mechanism of an enzyme[69]. The QM/MM approach regained interest in 1990 thanks to application and validation work by Field, Bash and Karplus and the basic QM/MM application to chemical problems[70]. Nowadays, QM/MM approaches have proven useful to accurately study reaction mechanisms in a protein environment using molecular dynamics and for ligand-protein binding studies[71,72,46,47,39].

The idea behind QM/MM methods[71] is that a small part of the system, usually containing the ligand and/or the binding site, is treated at the QM level, whereas the rest of the protein and solvent are treated at the MM level. In this way, the system is divided into a small QM part and large MM part as shown in Figure 2.6.

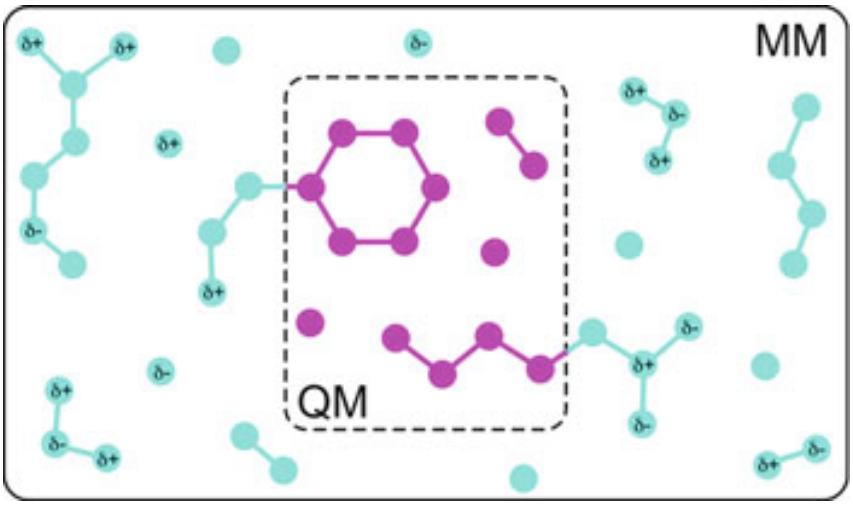

Figure 2.6 Schematic illustration of the QM/MM approach. (Image taken from ref [71])

The total QM/MM potential energy of the system includes three classes of interactions: between atoms in the QM region, between atoms in the MM region, and between QM and MM atoms. Interactions within the QM region are described at the QM level and those within the MM region are described at the MM level. Interactions between the two subsystems, on the other side, are more difficult to describe.

To study these types of interactions several approaches have been proposed, and they can be divided into two categories: subtractive and additive coupling schemes. In the subtractive scheme, different parts of the system undergo independent calculations at different levels of theory. The QM/MM energy of the total system is given by the

energy of the QM subsystem, calculated at the QM level, plus the energy of the complete system, evaluated at MM level, minus the energy of the QM subsystem, evaluated at the MM level.

$$E_{tot} = E_{QM}(QM) + E_{MM}(QM + MM) - E_{MM}(QM)$$

The last term is subtracted to correct for double counting of the contribution of the QM subsystem to the total energy. The main advantage of this scheme is that no explicit QM/MM coupling terms are needed due to the coupling between subsystems is handled at the MM level of theory. On the other hand, a disadvantage is that a force field is required for the QM subsystem; it also need to be flexible enough to describe the effect of chemical changes when a reaction occurs. One additional drawback of this method is the lack of polarization of the QM electronic density by the MM environment.

Instead, the additive scheme which is the most adopted approach for QM/MM calculations, adds the explicit treatment of the interactions between the QM and MM subsystems. Indeed, its main advantage is that the energy calculation of the QM region can be performed directly in the presence of the classical environment so that the electronic density of the QM region is optimized in the external electrostatic field of the environment. In the additive QM/MM scheme, the total energy of the system is equal to the sum of QM energy terms, MM energy terms and QM/MM coupling term:

$$E_{tot} = E_{QM}(QM) + E_{MM}(MM) + E_{QM/MM}(QM + MM)$$

The last term, the interactions between the two subsystems, can be treated with several available models: mechanical embedding, electrostatic embedding, and polarized embedding. They differ in the degree of polarization between the QM and MM regions.

- Mechanical embedding is equivalent to the QM/MM subtractive scheme outlined above, as it deals with QM/MM electrostatic interactions at the MM level (typically between rigid atomic charges). With this model both the QM and MM regions are not polarized.

- Electrostatic embedding allows for the polarization of the QM region since the QM calculation is performed in the presence of the MM charge model, typically by including the MM point charges as one-electron terms in the QM Hamiltonian. Therefore, the electronic structure of the QM subsystem can be adapted to the environment and the resulting QM density should be much closer to reality than that obtained from the mechanical embedding.
- In the polarization embedding scheme, both regions can polarize each other. Thus, not only the QM region is polarized by MM atoms, but the QM region can also induce polarization in the MM system. To obtain the total QM/MM energy, the MM polarizations need to be calculated at each step of the iteration of the self-consistent field of the QM wave function. Since the polarization is also calculated in a self-consistent manner, the QM/MM calculation can become very expensive.

The boundary between the QM and MM regions must be chosen meticulously, because if the QM and MM subsystems are connected by chemical bonds, it is necessary to be careful when making a cut. Covalent bonds often end up being cut by the QM/MM boundary, especially for systems such as proteins, but direct cutting of the bond will cause one or more unpaired electrons in the QM subsystem. However, these electrons are paired in bonding orbitals with the electrons that bind to the atom on the MM region. There are several approaches to overcome this artefact and saturate, or cap, the bond: link atom, capping potential or hybrid orbital.

- The link atom approach is the most widely used. It is based on replacing the MM part of the bond by an atom (link atom), usually a hydrogen atom, that is not used in QM/MM forces calculation and is not propagated during the MD simulation. In principle, each link atom provides three additional degrees of freedom to the system. In practice, the link atom is put in a fixed position along the bond at each step of the simulation, in order to remove these additional degrees of freedom. At each step, the force acting on the link atom is distributed over the QM and MM atoms of the bond.

- An alternative to the link atom approach is to replace a chemical bond between the QM and MM subsystem with a double occupied molecular orbital. This can be achieved either by the hybrid localized orbitals method, which introduces orbitals to the QM atom, or by the hybrid generalized orbitals approach, which places additional orbitals on the MM atom.
- The last approach that can be used to deal with the boundary between the QM and MM region is to replace the MM atom with a specifically designed pseudopotential[73] that mimics the real atom. In this approach the capping atom is included in the MD providing a more complete description of the system.

When performing a QM/MM simulation the classical MD part is driven by parameterized potentials (force fields), whereas in the QM region it is necessary to select a method to calculate the forces acting on the nuclei from the electronic structure calculations.

Given the non-relativistic Hamiltonian describing the many-body system consisting on interacting electrons and nuclei:

$$\hat{H} = \hat{T}_N + \hat{T}_e + \hat{V}_{Ne} + \hat{V}_{ee} + \hat{V}_{NN}$$

(where $\hat{T}_N$ and , $\hat{T}_e$ are the operators of the kinetic energy of nuclei and electrons, respectively, $\hat{V}_{Ne}$ is the electron-nuclei attractive potential, $\hat{V}_{ee}$ the electron-electron potential and $\hat{V}_{NN}$ the inter-nuclear repulsion potential), the application of the Born-Oppenheimer approximation allows to decouple the motion of nuclei from that of electrons. This allows to split the many-body wavefunction into the electronic and nuclear parts:

$$\Psi = \Psi_e \cdot \Psi_n$$

With this approximation, nuclei can be described as moving on the potential energy surface (PES) defined by the electronic potential. Therefore, it is possible to apply the same algorithms to propagate the system in time as discussed in classical MD. Usually, within the Born-Oppenheimer-based MD (BOMD)[72] one first solves the time-independent electronic Schrodinger equation for the electronic wave function ψ_e

$$\widehat{H}_e(r,R)\,\psi_e(\mathrm{r},\mathrm{R}) = E_e\,\psi_e(\mathrm{r},\mathrm{R})$$

where

$$\widehat{H}_e = \hat{T}_e + \hat{V}_{Ne} + \hat{V}_{ee}$$

and time dependence is injected through the parametric dependence on R, i.e. the classical dynamics of nuclei. Then, one uses ψ_e to get the instantaneous potential energy for the nuclei and to evolve the nuclear position as mentioned above. Forces acting on the nuclei are defined as the gradient of the expectation value of the electronic Hamiltonian with respect to the nuclear positions. The dynamics can be generated by integrating the Hamilton's or Newton's equations of motion.

A number of methods were derived to obtain the approximate solution to the electronic Schrodinger equation. The choice of a suitable QM method for performing QM/MM MD follows the same criteria as for pure QM studies, i.e. the ratio of accuracy and reliability to computational cost. Traditionally, semi-empirical QM methods have been the most popular, and remain important for QM/MM molecular dynamics simulations. Nowadays, with the increase in computational power, ab initio Hartree-Fock or density functional theory (DFT) are being widely used[72,46]. The applicability of HF method to biochemical problems is rather limited due to the lack of the electronic correlation; post-HF approaches, that mitigate this issue, are in most cases computationally expensive and this prevents their usage for significantly large systems such as those that are studied in biophysics. On the contrary, the DFT theory takes into account electronic correlation effects at the cost of Hartree-Fock calculations. DFT is built around the Hohenberg-Kohn theorem which states that the electron density of the non-degenerate ground state makes it possible to univocally determine all the properties of the ground state. Hence any observable physical quantity of the ground state, such as the total energy, can be expressed as a functional of the electron density. The energy functional is minimal when the density is a true ground state density. Using the variational approach, minimization of the energy functional leads to single-particle Schrodinger equations, called Kohn-Sham equations (the formalism based on them is thus called Kohn-Sham DFT (KS-DFT)). The solution of these equations yields a set of

KS orbitals which are not real orbitals of the physical system, but they can be used to compute the ground state density of the system. KS equations cannot be solved analytically, but numerically, in a self-consistent way. In the implementation of DFT codes, KS orbitals can be expanded as a linear combination of a finite basis set. In contrast to localized basis sets, such as Gaussian, the use of a plane waves basis set allows to greatly simplify the computations. The main disadvantage of plane waves is the need of a large amount of basis functions. However, the problem can be alleviated by the introduction of pseudopotentials replacing the explicit treatment of core electrons. Exclusion of core electrons reduces the number of orbitals to expand; in addition, valence orbitals are smoother compared to core ones, which reduces the minimal required size of the basis set.

Alternatively, ab initio MD can be performed without solving KS equations explicitly during each time step. A computational strategy to do this was proposed by Car and Parrinello (CP-MD)[74,72]. Within this approach, a quantum-classical problem (electron-nuclear system in BO approximation) is reformulated in terms of a two-component purely classical system. This is done by providing electrons with a fictitious mass. This choice allows to propagate the whole two-component classical system using a chosen classical integrator. After an initial standard electronic minimization, as done in BOMD, the fictitious dynamics of the electrons keeps them on the electronic ground state corresponding to each new ionic configuration visited during the MD, thus yielding accurate ionic forces. Due to the much smaller mass of electrons one needs to choose much smaller time step for integration (typically of around 0.12 fs) than the one usually employed in the force field-bases (usually 2 fs) or BOMD (usually 0.48 fs).

Nowadays, QM/MM simulations can be performed using many different programs, there are quantum chemistry codes that have incorporated some molecular mechanics features or the contrary. Moreover, interfaces that links two different software (one from molecular mechanics and one from quantum chemistry) have been developed. The QM software takes care of calculating the properties related to the QM region while the MM software is responsible for computing the MM properties and usually conducts the molecular dynamics simulation. The interface module ensures the

transfer of data between the QM and MM programs and usually compute the hybrid interaction between the MM and QM subsystem. This makes it possible to carry out hybrid QM/MM simulations even of large and complex systems, overcoming the limitations associated with using the two levels of theory separately.

METADYNAMICS-BASED APPROACHES FOR MODELING THE HYPOXIA-INDUCIBLE FACTOR 2α LIGAND BINDING PROCESS

3.1 Introduction

As discussed in Chapter 1 - Introduction, recently MD simulations are being used increasingly in the study of processes happening on timescales that range from nanoseconds to milliseconds and beyond,[75] making them attractive for the study of ligand binding. However, computation of key thermodynamic quantities requires the observation of multiple binding events to obtain reliable statistics on the process, thus increasing the computation time. Therefore, enhanced sampling techniques are used to speed-up the simulation of the binding/unbinding events[20,76]. Most of these techniques, presented in Chapter 2 - Methods, make use of a bias potential that forces the system to sample higher energy regions, speeding up the crossing of energy barriers.

Among the methods for studying ligand binding based on enhanced sampling MD[7,33,77–81], in this work we focused on steered MD[82] (sMD) and Metadynamics[25] (MetaD). SMD was inspired by single-molecule pulling experiments and applies a moving restraint bias that pulls the system along a selected variable. Despite its wide applications to the study of (un)folding mechanisms of proteins[83,84] and transportation of ions and other molecules across membrane channels[85,86], sMD has also emerged as a method for studying ligand (un)binding[22,58,87,88], given that it is particularly well designed for the investigation of entry and exit pathways. Its points of strength are the easy setup and the shortness of simulations[14]. On the other hand, sMD still suffers from several limitations, in particular regarding the calculation of the potential of mean force (PMF)[14]. As already discussed in the Chapter 2, during the pulling, a part of the work is spent as dissipative work and convergence can be difficult to reach. In theory, the Jarzynski's equality may account for the dissipative part of the work; however, when the range of work obtained in multiple replicas is broad, simulations with the lowest work contribute most to the calculation of the average work[89]. These limitations may be overcome by performing a large number of replicas and reducing the pulling speed, but for some complex systems this is often not enough.

As presented in Chapter 2, MetaD is a method based on the introduction of a history-dependent bias potential applied to a small number of suitably-chosen collective variables (CVs)[25,90,91]. The choice of the CVs is the most critical aspect in MetaD, and results can be seriously affected by the omission of important degrees of freedom (hysteresis).[90] Given that the computational cost to reconstruct the free-energy surface exponentially grows with the number of CVs, Branduardi et al.[28] developed the Path Collective Variable (PCVs) method, which allows exploration of complex multidimensional processes along a predefined pathway described by a single CV. An additional CV, that describes the distance from the reference path, usually completes the set of CVs necessary to efficiently sample the process of interest.

Here, we investigate the ligand binding process to the Hypoxia Inducible Factor 2α (HIF-2α), a pharmaceutically relevant system widely recognized as a target for cancer therapy[92]. HIF-2α mediates the physiological responses to hypoxia through

heterodimerization with the Aryl hydrocarbon Receptor Nuclear Translocator (ARNT)[93,94]. Under normoxia conditions (adequate oxygen levels), HIF-2α is hydroxylated by PDH (propyl-4-hydrolase), recognized by a second protein (VHL, von Hippel-Lindau) and finally degraded by the ubiquitin system (regulatory protein required for protein degradation in the ubiquitination process). In contrast, under hypoxic, oxygen-deficient conditions, hydroxylation cannot occur and HIF-2α accumulates in the cytoplasm; it is then transported into the nucleus where it dimerizes with ARNT to form the transcriptionally active heterodimer[95] (Figure 3.1).

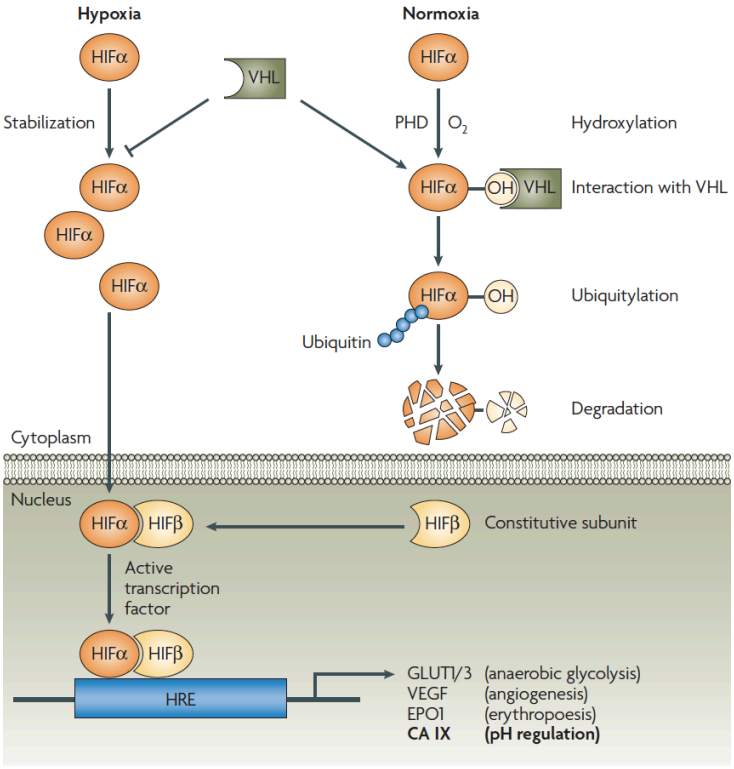

Figure 3.1 Mechanism of hypoxia-induced gene expression mediated by the HIF transcription factor. (Image from ref[95]).

Activated genes are involved in glycolysis, erythropoiesis and angiogenesis; the gene products include erythropoietin, that stimulates the production of red blood cells, and vascular endothelial growth factor (VEGF), a regulator of blood vessel growth[94]. In tumor masses, the abnormal vasculature creates hypoxic regions that activate HIFs to promote angiogenesis and to switch to anaerobic metabolism, sustaining cell viability under hypoxic conditions[96]. For this reason, targeting the HIF-2α:ARNT

interface with ligands has gained increasing attention as a therapeutical anticancer strategy.

Both HIF-2α and ARNT belong to the mammalian basic helix–loop–helix-PER-ARNT-SIM (bHLH-PAS) family of proteins, which members modulate transcriptional responses to environmental and cellular signals and are involved in a variety of physiological processes and diseases in humans[92,97]. Members of the bHLH-PAS family present an N-terminal bHLH region for DNA binding, two PAS domains (PAS-A and PAS-B) with the role of both sensing external signals and recognize the dimerization partner and a transactivation domain. For a long time, only another bHLH-PAS protein, the Aryl hydrocarbon Receptor (AhR), was known to be activated by binding to a wide range of ligands within its PAS-B cavity[98,99]. More recently, following the discovery of a buried cavity within the HIF-2α PAS-B domain[100], several artificial small molecules were identified as HIF-2α ligands and potential inhibitors of the HIF-2α:ARNT dimerization[101–106]. The structural determination of the HIF-2α:ARNT dimer encompassing the whole bHLH-PAS region, in the unbound, DNA-bound, and inhibitor-bound forms[93], recently allowed us to investigate the inhibition mechanism of the 0X3 antagonist and to shed light on pharmacophoric features required for the development of new inhibitors.[107]

In this work, we combined sMD and PCVs MetaD simulations to investigate the binding process of two known ligands to the HIF-2α PAS-B domain (Figure 3.2).

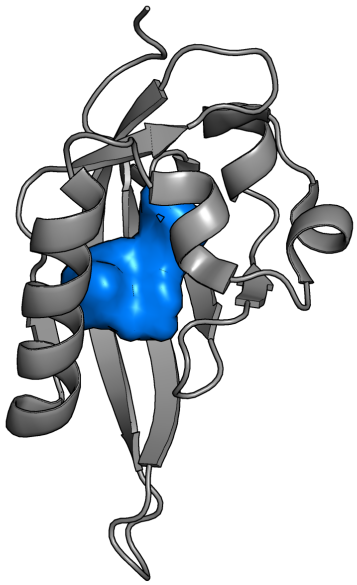

Figure 3.2 The 3D structure of the PAS-B domain of HIF2a is shown in grey and in cartoon, the ligand binding cavity is shown as surface in blue.

We were aimed at both investigating the ligand entrance pathway into the binding cavity and assessing the validity of the selected methods for such a complex system. In fact, the buried nature of the cavity makes it difficult to imagine the entry or exit route of the ligand and, despite a previous MD investigation identified probable pathways for water exchange with the bulk solvent,[103] the access of larger organic molecules to the cavity has never been studied. Moreover, it is conceivable that ligand entrance into this cavity may involve significant protein conformational rearrangements. The above features of the system make simulation of the ligand binding process a non-trivial task and required the development of specific methodological approaches. In the light of the obtained results, these methods appear to be suitable also for the elucidation of other ligand binding processes with similar characteristics.

The main results of the work reported is this Chapter were published in ref[108].

3.2 Methods

System preparation and molecular dynamics simulation

Crystal structures of HIF-2α in its bound state with the THS-020 ligand (PDB ID: 3H82[103]) were obtained from the Protein Data Bank (PDB)[109]. The PAS-B of the ARNT

protein partner, included in the X-ray deposition, was removed. This does not induce perturbations in the structure of the HIF-2α PAS-B, as shown by the RMSD plot (Figure 3.3) that highlights the stability of the HIF-2α domain during the MD simulation.

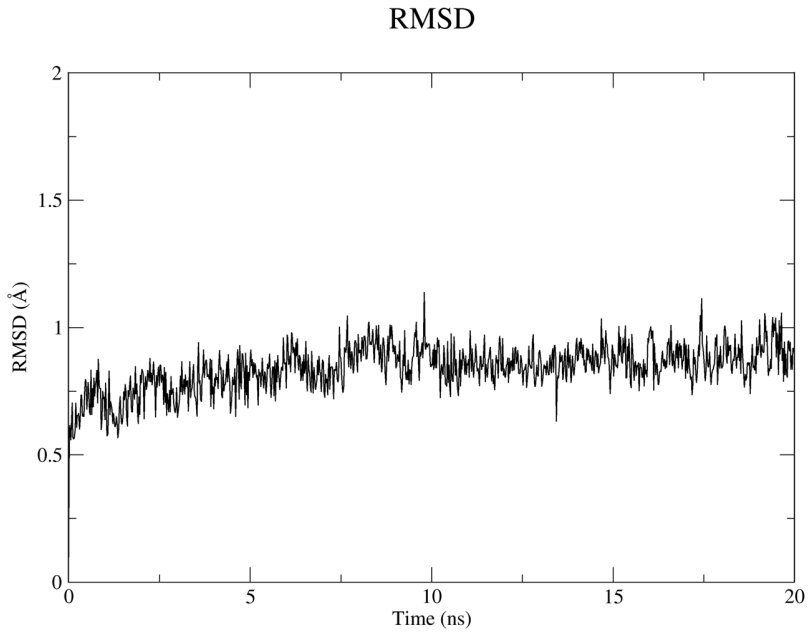

Figure 3.3 Plot of the RMSD values computed on Cα atoms for the HIF-2α PAS-B domain during the unbiased MD simulation.

The KG-721 bound form was obtained with molecular docking calculations (see the next sub-section). Protein was prepared with the Protein Preparation Wizard[110] included in Maestro: hydrogen atoms were added, all water molecules removed, C- and N-terminal cappings were added, disulphide bonds were assigned, and residue protonation states were determined by PROPKA[111] at pH = 7.0. The ligands were prepared using the LigPrep[112] tool included in Maestro in order to optimize the structures. The partial charges of ligands were calculated using the RESP[113] method at AM1-BCC[114] level of theory in Antechamber[115], while a GAFF[116] parametrization was used to achieve the complete topological description of each ligand. The unbiased MD simulations were performed using GROMACS 2018.6[117]. The protein was solvated in an orthorhombic box with TIP3P[118] water molecules, ad neutralized with Na^+/Cl^- ions. The minimal distance between the protein and the box boundaries was set to 20 Å. The Amber ff14SB force field[119] was used for the protein and a multistage equilibration protocol was applied: the system was first subjected to 2000 steps of steepest descent energy minimization, with positional restraints (239 kcal mol^{-1} nm^{-2}) for backbone and

ligand. Subsequently, a 200 ps NVT MD simulation was used to heat the system from 0 to 100 K with restraints lowered to 96 kcal mol^{-1} nm^{-2}; then the system was heated up to 300 K in 400 ps during an NPT simulation with further lowered restraints (48 kcal mol^{-1} nm^{-2}). Finally, the system was equilibrated during an NPT simulation for 2 ns with backbone restraints lowered to 12 kcal mol^{-1} nm^{-2}. In the NVT simulations temperature was controlled by the Berendsen thermostat[120] with coupling constant of 0.2 ps, while in the NPT simulations the V-rescale thermostat[121] (coupling constant of 0.1 ps) was used and the pressure was set to 1 bar with the Parrinello-Rahman barostat[122] (coupling constant of 2 ps). A time step of 2.0 fs was used, together with the LINCS[52] algorithm to constrain all the bonds. The particle-mesh-Ewald method[123] was used to treat the long-range electrostatic interactions with the cutoff distance set at 11 Å. Short-range repulsive and attractive dispersion interactions were simultaneously described by a Lennard-Jones potential, with a cut-off at 11 Å. Finally, a 20 ns production run was performed without the constraints.

Molecular docking of the KG-721 ligand

Conformational analysis of the ligand structure was performed using Macromodel[124] with the OPLS_2005[125] force field. The obtained global minimum was used as starting point for molecular docking calculations using Glide[126] XP[127] (Extra Precision). In particular, Glide uses a flexible ligand-rigid protein approach, in which a series of hierarchical filters are applied to find the possible positions and conformations of the ligand in the binding cavity (poses). The properties of the protein are represented on a grid of points on which the contributions that each protein atom gives to the interaction energy are pre-calculated and stored, in order to reduce computing time and cost, and to provide gradually more accurate scores. The initial screenings are deterministically performed over the complete phase space of the ligand to identify the most promising poses. From the selected poses, the ligand is then refined in the torsional space in the receptor field. To take into account the flexibility of the protein, the ensemble-docking approach was used, that involves ligand docking to multiple receptor conformations. These can be derived either experimentally or computationally (e.g. by MD simulations)[128]. The conformational ensemble here

selected consisted of the crystallographic structures of the HIF-2α PAS-B in complex with artificial ligands available in PDB (3F1O[100], 3H82[103], 3H7W[103], 4GS9[101], and 4GHI[102]). The results showed that the best XP score is the one related to the KG-721 ligand in the 4GHI structure.

Steered MD simulations (sMD)

All the sMD simulations were performed using the PLUMED 2.4.6[129,130] plugin integrated in GROMACS 2018.6[117]. We chose the ligand-protein distance as the pulling variable. This was defined as the distance between the center of mass of selected atoms at the bottom of the binding cavity (different for the 2 pathways, see Figure 3.4) and the center of mass of the ligand heavy atoms.

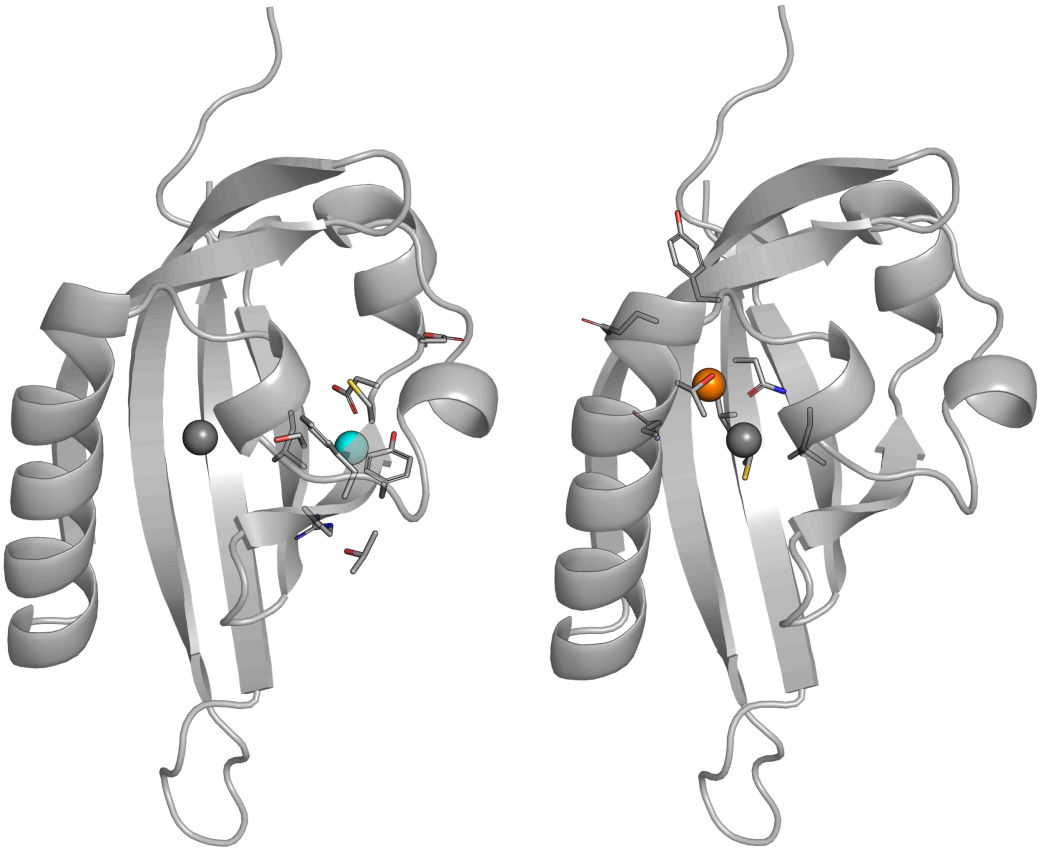

Figure 3.4 Representation of the pulling variable for the two paths (path 1 on the left, path 2 on the right). In cartoon, the PAS-B domain of HIF2a; in gray sphere, the center of mass of the heavy atoms of the ligand (THS-020); in cyan (orange) sphere, the center of mass of N, Cα, C and O atoms of selected reference residues: for path 1, L245, S246, R247, F254, T255, Y256, C257, D258, D259; for path 2: L245, E320, T321, Q322, G323, C339, V340, N341, Y342.

The spring constant was set to the value of 10.0 kcal/mol·Å^2 and the ligand was pulled from the initial value of CV to 35 Å in 25 ns with a resulting pulling velocity of

0.984 Å/ns. We ran 50 independent replicas and the time length for each simulation was 25 ns, which ensured the achievement of a complete solvation of the ligand in the unbound state. The starting point of each replica was derived from an ensemble of states extrapolated at regular time intervals of 0.2 ns from the last 10 ns of the unbiased simulation.

Metadynamics (MetaD) and Path Collective Variables (PCVs)

The Metadynamics method with PCVs formalism has been widely used to investigate biological processes, to compute their free-energy surfaces, and characterize their kinetic behavior[131,132]. In this work, PCVs were used to study the transition between the bound and the unbound states in the unbinding process of some HIF2-α ligands. The reference path for MetaD-PCVs simulations was created for each ligand starting from the sMD simulation with the lowest value of the unbinding work. Using an in-house developed script implemented in VMD[133], the RMSD matrix was calculated for a selection of protein atoms (Figure 3.5) and all the ligand heavy atoms.

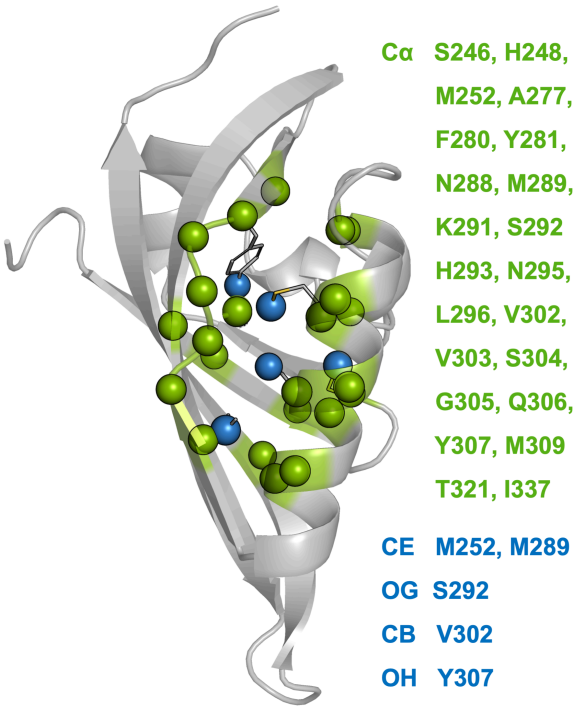

Figure 3.5 Selection of protein atoms for RMSD calculations. In green spheres, the Ca atoms; in blue spheres, some sidechain atoms.

The optimal reference path should have a regular symmetric matrix (with a typical gull-wing shape)[28]. Anyway, in sMD simulations, when the ligand reaches the unbound

state, it freely moves in the solvent thus providing an inhomogeneous frame-to-frame distance, and an irregular matrix. To avoid this problem, only the frames belonging to the first part of the path (ligand in contact with the protein) were extracted from the sMD simulations. Starting from the first frame (F, ligand in the bound state), equally spaced frames with a distance of 2 Å from the previous one were selected. In the last frame of this part (M) the ligand is located near to the mouth of the cavity. The ligand was then translated in the bulk solvent (frame L) and the second part of the path was obtained by a 2 Å linear interpolation between frames M and L. The combination of the frames selected from the sMD simulation and those obtained with linear interpolation provided the reference path. In particular, 12 frames were used for the THS-020 ligand and 11 frames for the KG-721 ligand. For the first part of the path, the frames (from 1 to 7 for THS-020 and from 1 to 6 for KG-721) were obtained from the sMD simulations. For the second part of the path, the frames were obtained with linear interpolation, as described above. Following the procedure proposed by Branduardi et al.,[28] we introduced the two collective variables: s(R), the progress along the reference path; and z(R), the distance orthogonal to the reference path. The λ value was set to 33.0 nm^2. The distance between the instantaneous conformational state during the simulation and the reference coordinates in the path was evaluated by the RMSD metric[68]. In all simulations, the Gaussian-shaped potentials were deposited every 500 simulation steps, the initial height was set to 1 kJ/mol and the decay corresponding to a bias factor of 10 was chosen. The Gaussian widths (σ) for the s(R) and z(R) variables were set to 0.05 and 0.007, respectively. Widths were set so that they are about 1/3 of the CVs standard deviations observed in the unbiased MD simulation. The two variables, s(R) and z(R), were constrained to be less than 12 and 0.2 nm^2 respectively.

Extraction of minima and cluster analysis

To characterize the different minima identified on the final free-energy surface (FES), we extracted a group of frames belonging to each minima hole. To obtain a representative structure of the complex in each minimum, a cluster analysis on the metadynamics trajectory frames with a stride of 10 ps was performed. The GROMOS[134] clustering algorithm was applied, with a 2 Å RMSD cut-off on the heavy atoms of the

ligands. The centroid of the most populated cluster was then defined as the representative structure in that minimum.

3.3 Results

The analysis of the unbinding pathways was performed for two of the HIF-2α ligands identified in the study of Key et al[103]. The THS-020 ligand (Figure 3.6A) has a good binding affinity for the protein (ΔG_{exp} = -7.9 ± 0.5 kcal/mol) and the ligand-protein bound structure, determined by X-ray crystallography[103], is available. The KG-721 ligand (Figure 3.6B) is a less affine ligand (ΔG_{exp} = -6.9 ± 0.1 kcal/mol) identified in the same work[103]. We choose it among the other HIF-2α ligands[100–103,135,104–106,136,137], not only for the different binding affinity, but also to deal with a molecule not congeneric to THS-020[103,101], with different physico-chemical properties and with lower size. Moreover, for this ligand no experimental structures of the ligand-protein complex are available, thus offering us the opportunity to study a system where the starting conformation, obtained by docking, could not take into account the induced fit effects on the protein.

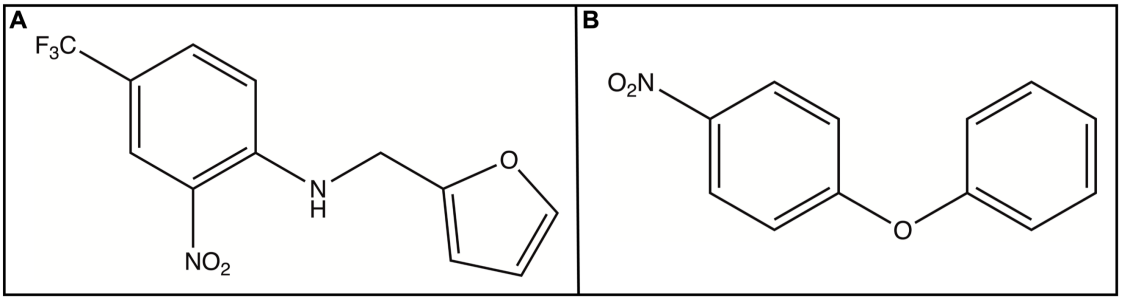

Figure 3.6 The 2D structure of THS-020 (A) and KG_721 (B).

In the following two sub-sections we present the application of a specific sMD-MetaD protocol on the THS-020. We first used sMD to identify the unbinding pathway, and then we applied PCV-MetaD simulations to characterize the relevant states in the binding/unbinding process and to obtain a reliable estimate of the binding affinity. In the third sub-section we present the results obtained with the same protocol for KG-721.

Identification of unbinding pathways for the THS-020 ligand by the sMD method.

Starting from the X-ray structure for the HIF-2α PAS-B domain in complex with THS-020, the possible ligand unbinding pathways were investigated using the sMD approach. Steered molecular dynamics is a popular method for studying ligand-protein unbinding events[138–140] and can provide both a qualitative description of the pathways and a quantitative estimate of the free-energy difference between the bound and unbound states. To calculate the free-energy difference using the Jarzynski equality it is necessary to have a high number of sMD replica. To this aim, a 20 ns unbiased MD simulation was performed starting from the X-ray structure, generating an ensemble of 50 slightly different states of the complex (Figure 3.7a), extracted from the last 10 ns. These were then used as starting points for the sMD simulations. The structural convergence of the unbiased simulation was assessed by calculating the RMSD matrix on the protein Cα atoms and on the ligand heavy atoms (Figure 3.7b).

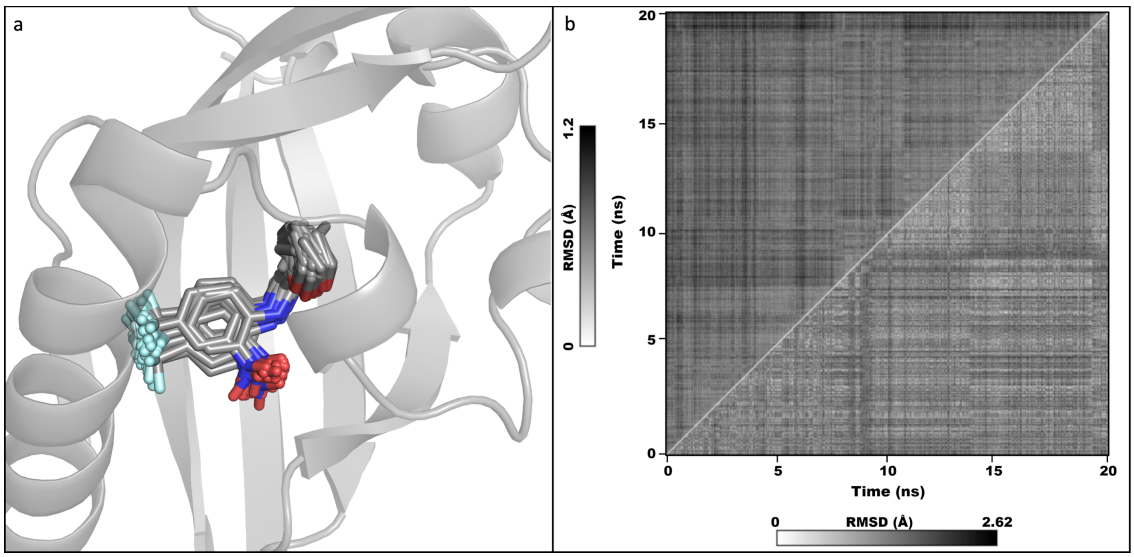

Figure 3.7 Monitoring of the conformational changes during unbiased MD simulation of the HIF-2α PAS-B with the THS-020 ligand. a) Representation of the ensemble of 50 ligand conformations in the last 10 ns of simulation. The protein structure is represented as cartoons and the different states of the ligand as sticks. b) RMSD matrix computed on Cα atoms (upper half) and on ligand heavy atoms (lower half).

Other authors identified two entry/exit pathways for solvent water by MD simulations of the apo HIF-2α PAS-B[103]: path 1 get through the Fα helix and the Gβ strand, while path 2 through the Fα helix, the short Eα helix and the AB loop (Figure 3.8). On this basis, for each of these two pathways we calculated a CV allowing to pull

the ligand out of the binding cavity, by selecting an appropriate set of residues at the bottom of the binding cavity (see paragraph 3.2 Methods).

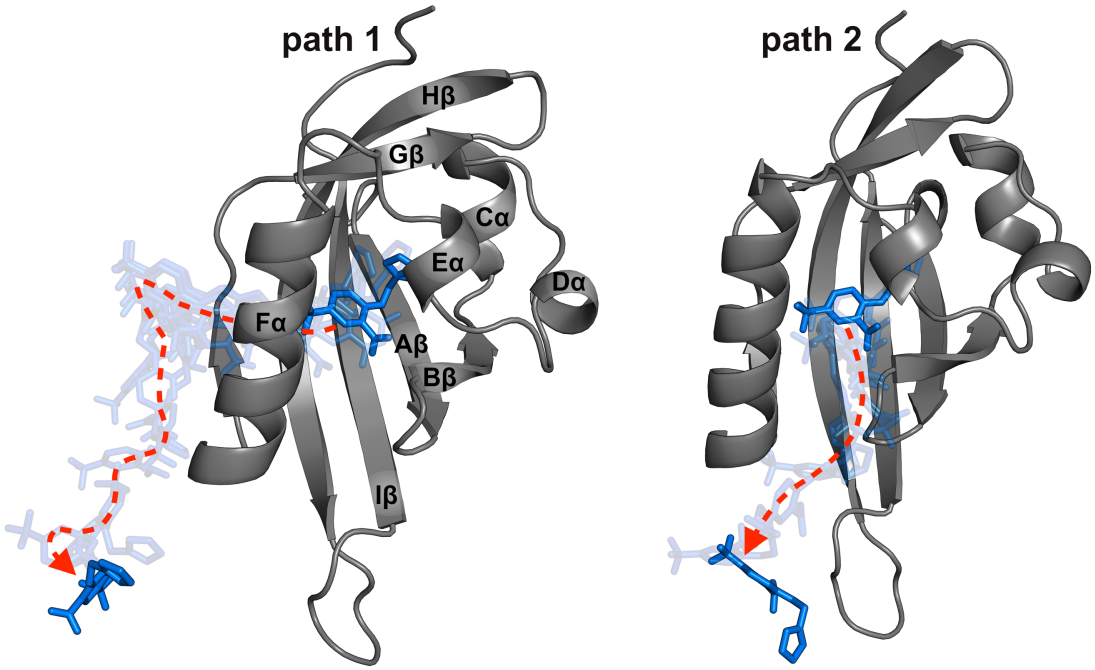

Figure 3.8 THS-020 unbinding pathways. In pathway 1 (left) the ligand passes through Fa and Gβ, while in pathway 2 (right) through Fa, Ea, and the AB loop. The starting protein structure is represented as grey cartoons, the ligand conformations in the first and last frames of the of the trajectory as blue sticks, and the conformations of the ligand in the intermediate frames as transparent sticks.

Before running the sMD simulations, it was necessary to define optimal values of various parameters such as the spring constant, the pulling velocity and, consequently, the time length for each simulation. As a first step, several spring constant values were tested: 0.1, 1.0 and 10.0 kcal/mol· Å^2. The results showed that the lowest values (0.1 and 1.0 kcal/mol· Å^2) were not sufficient to achieve the unbinding of the ligand since it remains significantly behind the bias position. For this reason, the chosen value for the spring constant was 10.0 kcal/mol· Å^2.

The choice of pulling velocity value is related to the time length of the simulation. We decided to perform tests on simulations of reasonable computational cost for this phase of the protocol, and to adapt the pulling speed accordingly. We compared 50 replicas of 5 ns, 50 replicas of 25 ns and 20 replicas of 200 ns. Given that the initial position of the CV is 10.4 Å and we set the end of the simulation when the ligand reaches the distance of 35 Å, the pulling velocity for the three sets of simulations were

set, respectively, to: 4.92 Å/ns, 0.984 Å/ns and 0.123 Å/ns. All simulations were performed either for pathway 1 or pathway 2. The work profiles resulting from the different replicas (in Figure 3.9) consistently show an increase of the work value during the initial part of the unbinding process followed by relatively settled work values, indicating the absence of interaction between the ligand and the protein. All the curves show a similar profile, but a qualitative comparison of the total work reveals that less work is required for unbinding following pathway 1. Moreover, a broader range of values is observed for the replicas following pathway 2, indicating that higher barriers can occur in some of the replicas associated to this pathway.

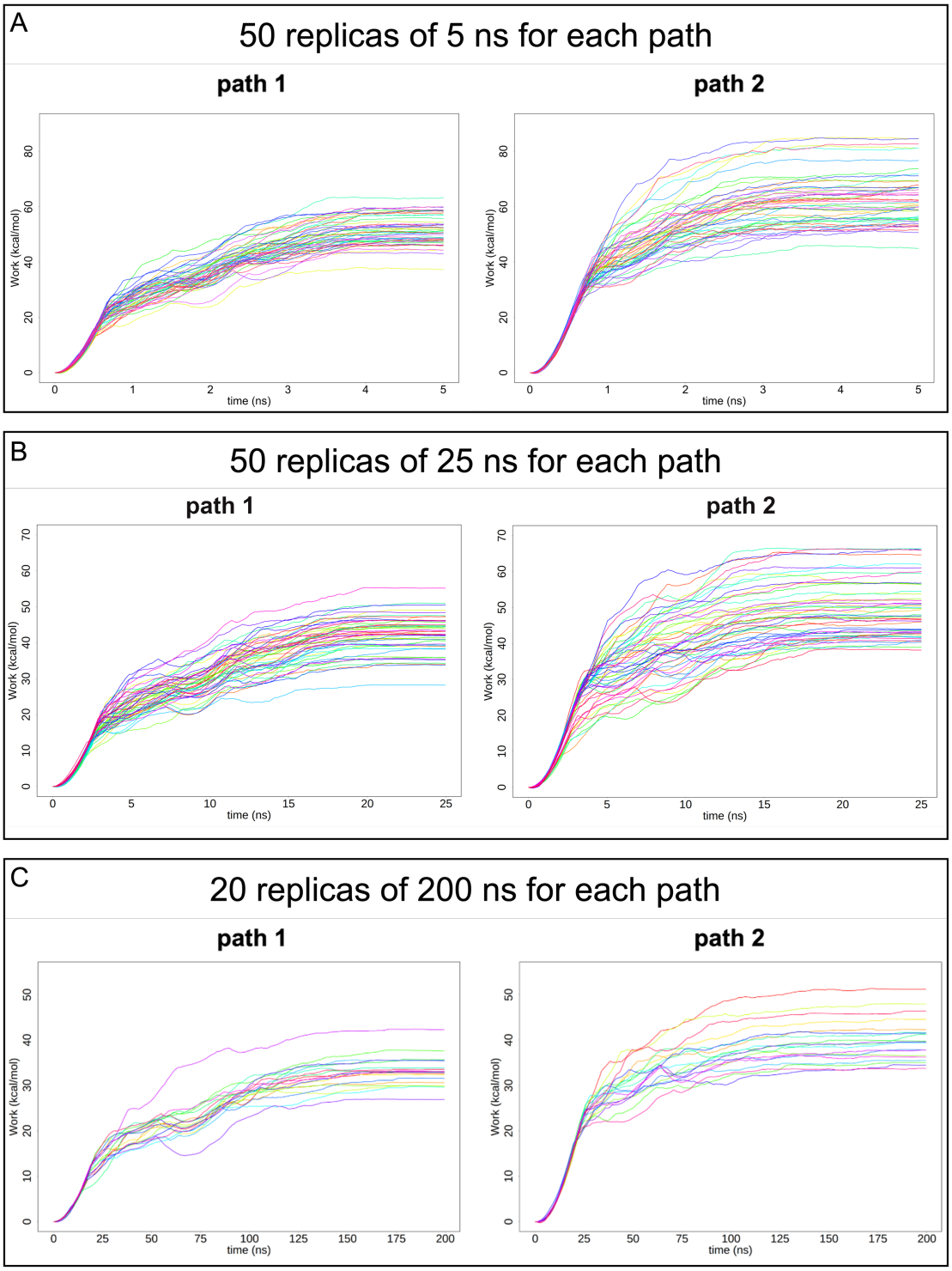

Figure 3.9 Work profiles for the two unbinding pathways of THS-020 obtained from sMD simulation with different time length: A) 50 replicas of 5 ns, B) 50 replicas of 25 ns, C) 20 replicas of 200 ns. The 50 curves of the work exerted on the system to pull the ligand along path 1, on the left, and along path 2, on the right.

A preliminary comparative analysis on the effects of using different pulling speed can be achieved by calculating the value of the minimum work required to pull the ligand outside the binding cavity (Table 3.1).

Table 3.1 Comparison of miminum work values obtained from sMD simulations with different time length.

	sMD of 5 ns	**sMD of 25 ns**	**sMD of 200 ns**
Pathway	W_{min} (kcal/mol)	W_{min} (kcal/mol)	W_{min} (kcal/mol)
1	37.35	28.19	25.46
2	44.92	38.16	33.77

The results confirmed a clear preference of the pathway 1 over pathway 2 in all the cases tested. Furthermore, it is possible to highlight that the values obtained for the 5 ns replicas are higher than those obtained for the longer replicas. This suggests that the parameters set for the 5 ns replicas are not appropriate for describing the unbinding of the THS-020 ligand. On the other hand, comparing the values obtained from the 25 ns and 200 ns simulations, there are small differences, also considering the much higher computational effort required to complete the 200 ns replicas. In addition, a geometric comparison on the bound portion of the unbinding pathways sampled with sMD simulations of 25 ns and 200 ns was performed, by computing the RMSD values (on ligand heavy atoms) between pairs of frames belonging to replicas of pathway 1. All the trajectories were aligned to the same reference structure (on Cα atom) and sub-sampled obtaining trajectories of 500 frames (stride=50ps for 25ns replicas and stride=400ps in the 200ns replicas). The application of the stride to the simulations allows a comparison between the frames obtained from the 25 ns and 200 ns replicas because the sMD bias is always at the same value. The RMSD was then computed among all pairs of frames in the same position of the trajectory in the different replicas and the results are shown in Figure 3.10.

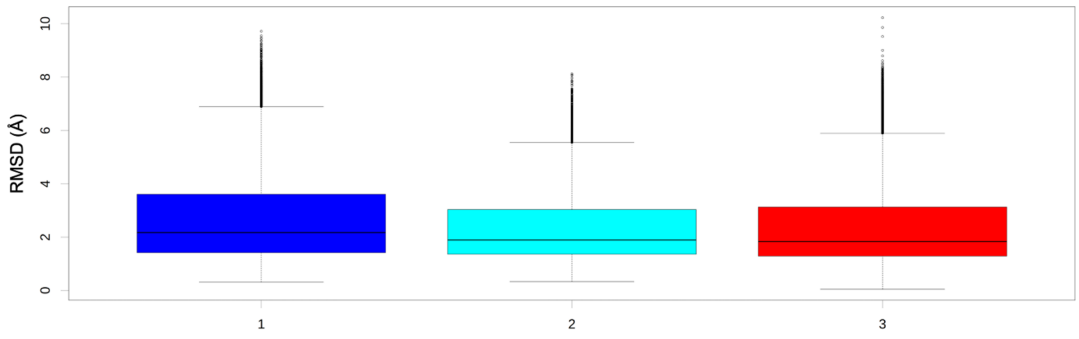

Figure 3.10 The boxplot of RMSD values between pairs of 25 ns replicas (blue boxplot), 200 ns replicas (cyan boxplot) and mixed 25 ns and 200 ns replicas (red boxplot). Computations of RMSD were performed on the first 40% (bound portion) of the sMD replicas.

The results show that the distribution of the RMSD values computed between frames coming from replicas at the same pulling speed (blue and cyan boxplots) is comparable to the one obtained comparing frames from replicas at different pulling speed (red boxplot). This means that the pathways sampled using a higher pulling speed are geometrically comparable to those produced at lower pulling speed. With the aim of employing the sMD to identify the preferred pathway and to use the results obtained as a basis for the construction of the reference path for the MetaD simulations, the 25 ns replicas were used, as they are less computationally demanding.

Further analyses were conducted in order to verify that the values chosen for the sMD simulation parameters were appropriate. In particular, the RMSD plot on Cα atoms, in Figure 3.11, and the secondary structure conservation graphs, in Figure 3.12, were calculated for the 50 replicas of 25 ns along path 1 and they revealed that no significant distortions of the protein structure (except for a slight deformation of Fα helix upon ligand unbinding) were observed during simulations. This is a further validation of the proposed protocol.

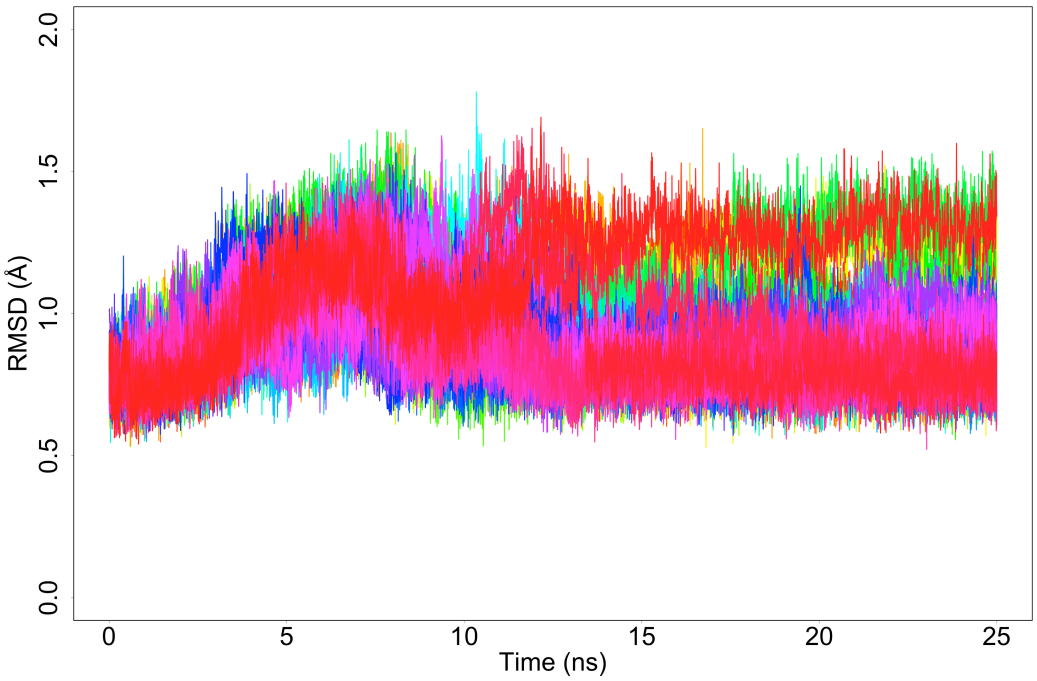

Figure 3.11 Plots of the RMSD values computed on Cα atoms for the THS-020 sMD replicas along path 1.

Figure 3.12 Plot of the secondary structure assignment for the HIF-2a PAS-B domain during the THS-020 sMD replicas along path 1. Secondary structures were assigned according to DSSP.

A more detailed analysis was performed for the 25 ns sMD replicas; the minimum and maximum work values (W_{min}, W_{max}), the minimum value of the maximal force (F_{max}) among the replicas, and the free-energy difference between the unbound and the bound states (ΔF_{unbind}) were extracted for each pathway (Table 3.2). The results show that the W_{min} necessary to pull the ligand outside the cavity along path 1 is about 10 kcal/mol less than that required for path 2; a similar trend is observed for the values of W_{max} and ΔF_{unbind}. A difference of 75.59 pN between the F_{max} values in the two paths is observed, which confirms a clear preference of path 1 over path 2.

Table 3.2 Results of sMD simulations for the THS-020 ligand.

Pathway	W_{min} (kcal/mol)	W_{max} (kcal/mol)	F_{max} (pN)	ΔF_{unbind} (kcal/mol)	st.dev.
1	28.19	55.17	1013.77	30.55	16.80
2	38.16	66.42	1089.36	40.25	13.10

Therefore, steered MD allowed us to compare the two unbinding pathways of THS-020 and to select the preferred one by identified a higher energy barrier along pathway 2. However, sMD provided a value for the ΔF_{unbind}, estimated by means of Jarzinsky equality, that was about 4 times higher than the experimental value. It is indeed known that these non-equilibrium simulations generally undersample the relevant protein−ligand states across the unbinding pathway, leading to errors in the computed binding free-energy[141]. Moreover, replicas with lower work done on the system have an enormous weight compared to all the other trajectories, which makes the method extremely sensitive to insufficient sampling[21]. For this reason, we then applied the PCVs MetaD approach that was recently proposed as a valuable method to compute absolute binding free-energies in ligand binding[132,131,142].

Metadynamics simulations and Free-Energy Profiles with Path-CVs for THS-020.

For a detailed mechanistic interpretation of the ligand binding/unbinding process, we used well-tempered metadynamics simulations with the PCVs approach[143,142,144]. This allowed us to characterize the relevant states along the preferred path obtained with sMD simulations (the one with lower values of total work obtained from the sMD simulation, path 1), as well as to estimate the binding free-energy value. The key points of this method are the choice of appropriate CVs and the construction of a set of equally spaced frames along the CVs in terms of RMSD between adjacent snapshots. This frameset represents a reference path for investigating the process. As CVs we used: the progress along the path, s(R); and the distance orthogonal to the reference path, z(R). We want to underline the importance of the path construction phase, especially in a case with a buried binding site like the one presented in this work. Here we decided to include both ligand and protein atoms in the frameset that represents the path to better describe the protein conformational changes during the process

(mouth opening through sidechain conformational changes and small backbone adjustment). Only protein atoms involved in the conformational changes, highlighted by sMD simulations, were included. Moreover, a hybrid approach that combines frames from sMD simulations and linear interpolation was used for the inner and outer parts of the path, respectively, as discussed in paragraph 3.2 Methods. The reference path obtained with this approach is represented in Figure 11. The RMSD matrix of the frameset (Figure 3.13) is a symmetric matrix with a typical gull-wings shape, indicating that the frames are correctly equally spaced.

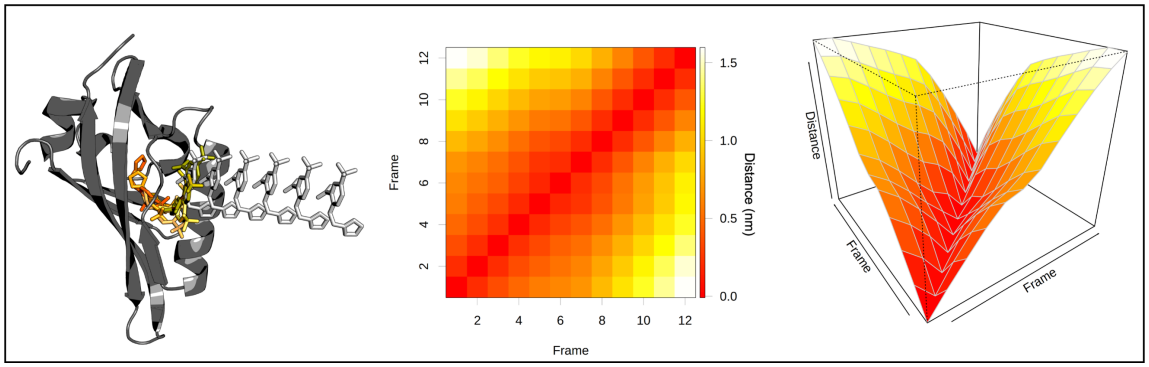

Figure 3.13 Resulting reference path for PCVs for the THS-020 ligand. Protein is represented in the bound conformation as dark grey cartoons, the ligand in the first part of the path (frames from sMD) as sticks from orange to olive, and the ligand in the second part of the path (frames extrapolated from linear interpolation) as light grey sticks. 2D (left) and 3D (right) representation of the RMSD matrix obtained from the frameset built for the THS-020 ligand.

We collected a total of 1.8 µs of metadynamics simulation in which we observed several binding/unbinding events, as shown in Figure 3.14A. The binding free-energy (ΔF_{bind}), calculated as the free-energy difference between the deepest minima in the bound state and the flat plateau in the unbound state, turns out to be equal to -11.8 kcal/mol. The free-energy profile during the simulation, shown in Figure 3.14B, indicates that the simulation reaches a constant value of ΔF_{bind} after about 1200 ns.

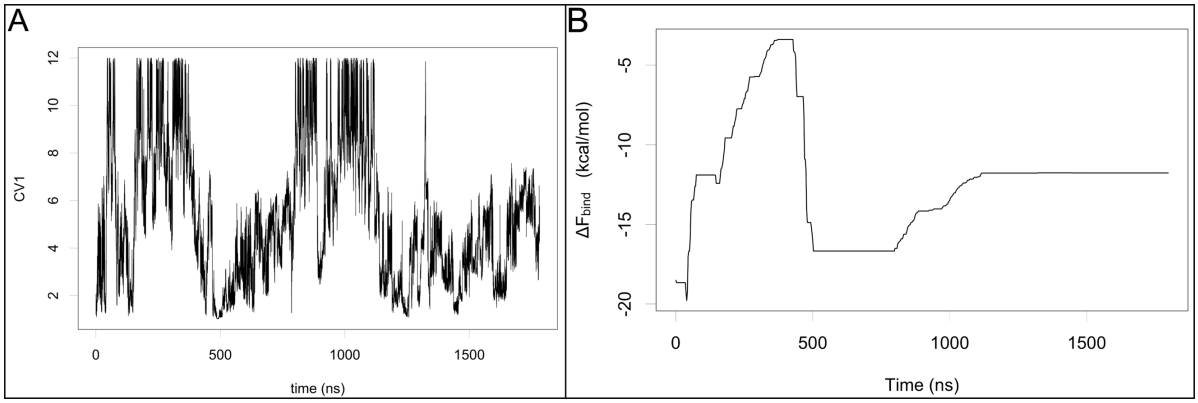

Figure 3.14 A) Plot of the CV1 (s(R)) against simulation time during THS-020 MetaD simulation: the lowest values of s(R) correspond to the bound state, while the highest to the unbound ones; B) one-dimensional projection of the binding free-energy values associated to the path 1 during the metadynamics simulation.

The convergence was also monitored by plotting the hill heights as a function of the simulation time, Figure 3.15.

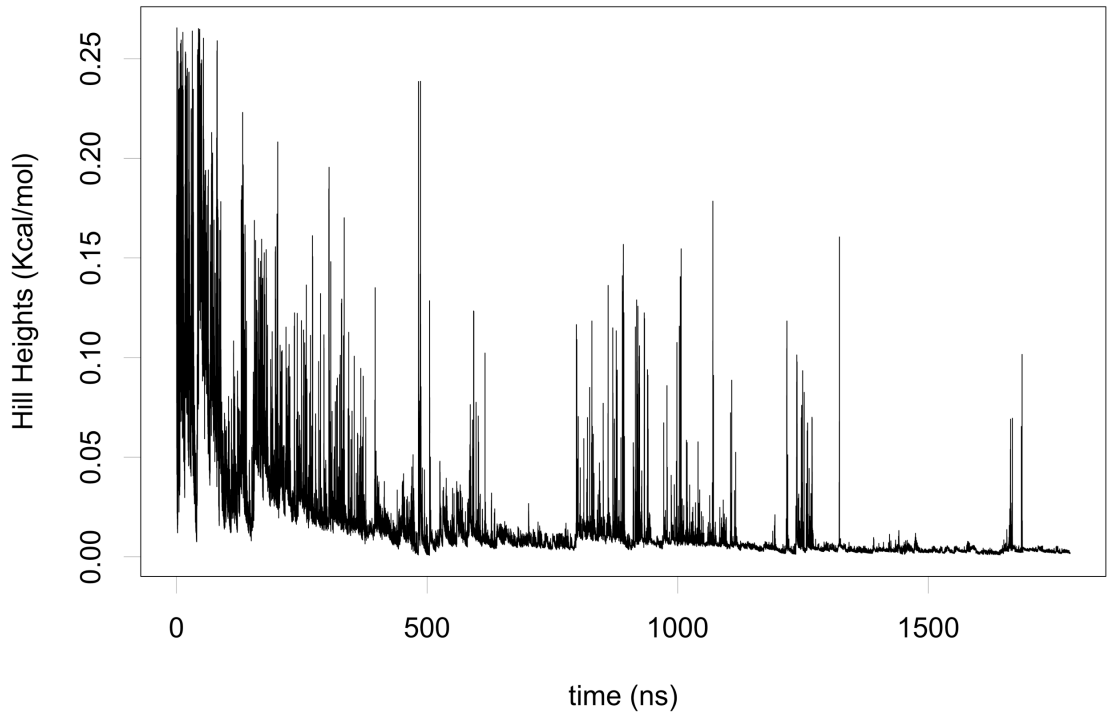

Figure 3.15 Plot of the hill heights during the metadynamics simulation for the THS-020 ligand.

The free-energy surface (FES) for the binding/unbinding process, as a function of CV1 (s(R)) and CV2 (z(R)) is shown in Figure 3.16 together with the relevant minima found along the pathway (labeled as A to G). The coordinates and the binding free-energy values of each minimum are reported in Table 3.3.

Table 3.3 Coordinates and binding free-energy values of the relevant minima along the CV1 (s(R)) and CV2 (z(R)) for the THS-020 ligand.

Minimum	s(R)	z(R)	ΔF (kcal/mol)
A	1.214 1.360	-0.003 0.017	0.00
B	2.691 3.019	0.042 0.056	5.65
C	1.922 2.146	0.124 0.136	2.50
D	3.006 3.209	0.088 0.111	5.78
E	4.872 5.583	0.136 0.165	4.57
F	5.262 5.411	0.024 0.029	6.91
G	6.694 6.772	0.084 0.098	7.63

A cluster analysis of the conformations belonging to each minimum hole was then performed. In each minimum, the centroid of the most populated cluster (the first one) was used as the representative structure for that minimum. Looking at the FES (Figure 3.16), three different regions can be identified following CV1: the bound state, with s(R) values between 1 and 3 (minima A-C); the intermediate states, with s(R) values between 3 and 7 (minima D-G); and the unbound state, with s(R) values from 7 onwards. In the deepest minimum (A), the ligand is oriented in the same way as in the X-ray structure (ligand RMSD=1.36 Å). This geometry is stabilized by a hydrogen-bond between the NH group of the ligand and the H248 residue as well as by a transient hydrogen-bond between the oxygen atom of furan and the S246 residue.

Moving up to higher CV2 values, alternative binding geometries can be detected. In the minimum B, the ligand is shifted towards the exit of the cavity and breaking of the hydrogen-bonds that stabilize minimum A causes a lower stability. Instead, in

minimum C the ligand is located on the mouth of the binding cavity and is even turned of 180°, with the CF_3 group oriented towards the bottom of the cavity. Also, in this minimum, the amino group of the ligand forms a hydrogen-bond, with the S292 residue. This last conformation represents the second minimum in energy.

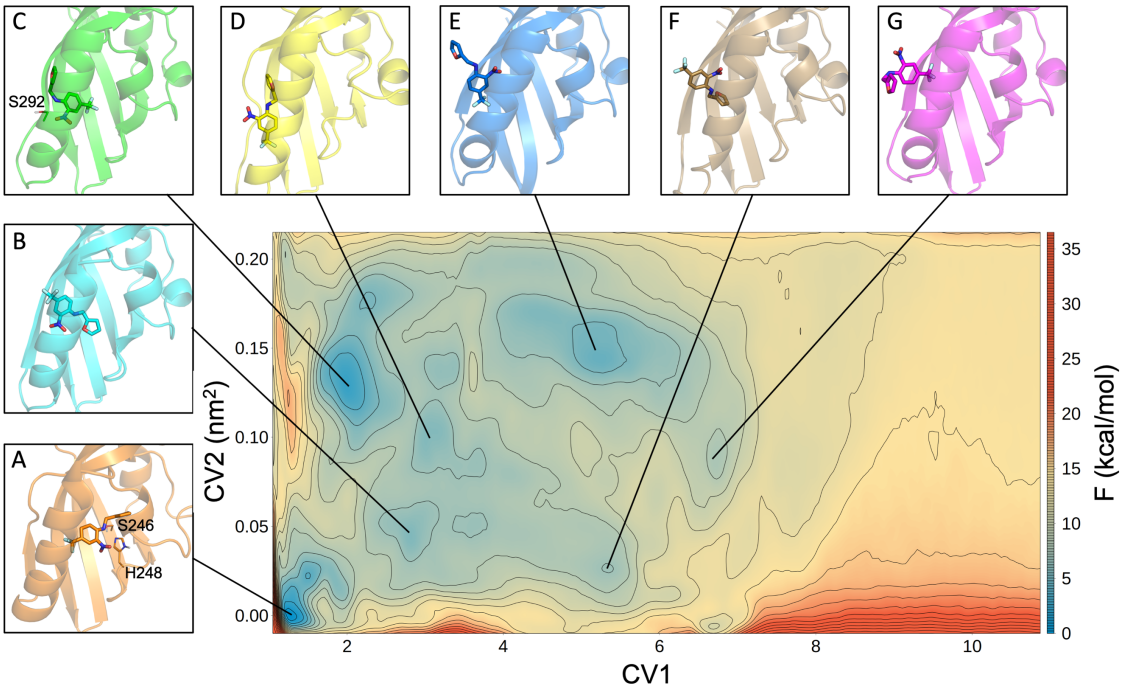

Figure 3.16 Free-energy surface obtained from the PCVs approach for the binding/unbinding of the THS-020 ligand. The isolines are drawn using 1.5 kcal/mol spacing. The 3D structures of the centroids of the main minima are reported with different colors: the protein is represented as cartoons and the ligand as sticks. The black lines indicate the corresponding minima in the FES.

Unbinding pathway for the KG-721 ligand

Following the encouraging results on THS-020, for which the selected methods were able to identify the experimental binding geometry as the most stable among all the possible bound states, we extended the study to the lower affinity KG-721 ligand. Given the lack of an experimental structure, we obtained the starting geometry for our calculations by molecular docking, using the ensemble-docking technique[128] (details are reported in paragraph 3.2 Methods). This technique has led to improve the description of ligand-induced protein conformational changes in many systems[128,145–147] but it may be not sufficient in some particularly challenging cases. These include

docking studies of ligands different from the ones co-crystallized in the protein structures of the ensemble (if any), like in our case.

As for the THS-020 ligand, 50 independent replicas of sMD simulations (each of 25 ns) were performed for the two possible pathways.

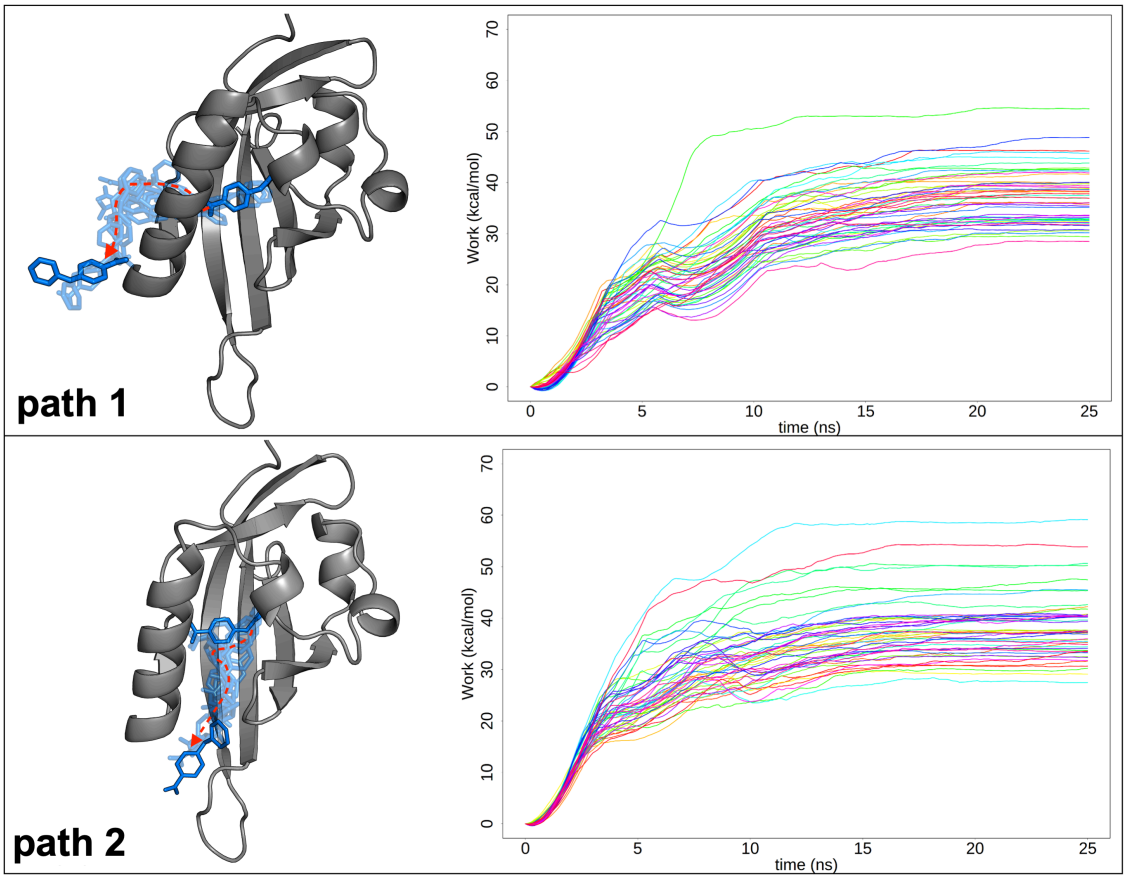

Figure 3.17 KG-721 unbinding pathways. In pathway 1 (above) the ligand passes through Fα and Gβ, while in pathway 2 (below) through Fα, Eα and the AB loop. The starting protein structure is represented as grey cartoons, the ligand conformations in the first and last frames of the trajectory as blue sticks, and the conformations of the ligand in the intermediate frames as transparent sticks. On the top left, the 2D structure of the ligand is reported. On the right, the work profiles for the two paths.

The resulting work profiles (Figure 3.17) are similar to those obtained for THS-020 But, at difference with that ligand, they show a similar range of work values for path 1 and path 2 and do not suggest any preference for one path over the other. This is also confirmed by negligible differences between the values of W_{min}, W_{max} and ΔF_{unbind} (Table 3.4) in the two paths. This result suggests that for a small ligand the two pathways may have a similar probability. This hypothesis is consistent with the findings of Key et al[103], which observed a similar percentage of transferring of the small water molecules in the two paths. However, the observed difference of 105.08 pN

between the F_{max} values of the two paths of KG-721 suggests that a higher barrier for unbinding exists along path 2, similarly to what was observed for the THS-020 ligand.

Table 3.4 Results of sMD simulations for the KG-721 ligand.

Pathway	W_{min} (kcal/mol)	W_{max} (kcal/mol)	F_{max} (pN)	ΔF_{unbind} (kcal/mol)	st.dev.
1	28,46	54,45	851,21	30,55	13,04
2	27,47	59,13	956,29	27,12	16,60

Based on the results obtained with the Steered MD simulations, 11 frames along path 1 were used to build the reference path for metadynamics simulations. The resulting RMSD matrix of the frameset and a representation of the reference path are shown in Figure 3.18.

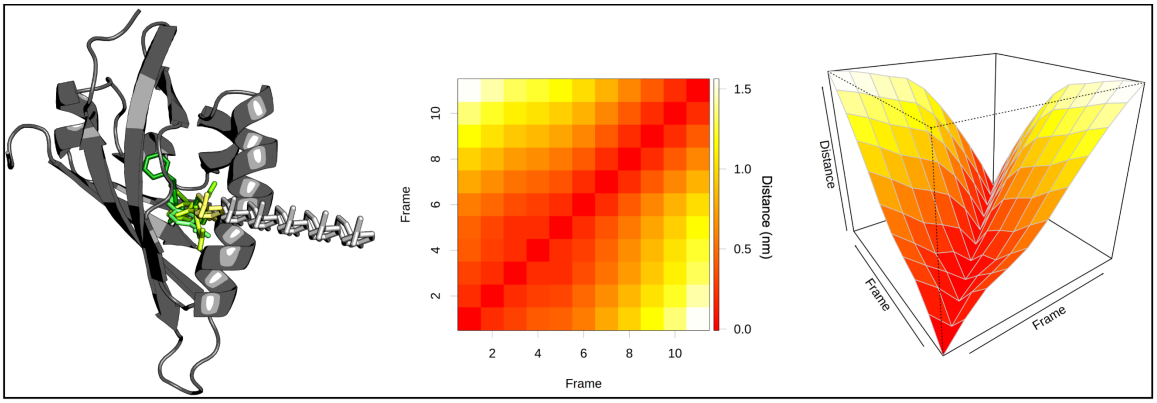

Figure 3.18 Resulting reference path for PCVs for the KG-721 ligand. Protein is represented in the bound conformation as dark grey cartoons, the ligand in the first part of the path (frames from sMD) as sticks from green to limon, and the ligand in the second part of the path (frames extrapolated from linear interpolation) as light grey sticks. 2D (left) and 3D (right) representation of the RMSD matrix obtained from the frameset built for the KG-721 ligand.

After 3 μs of metadynamics simulation, we reconstructed the free-energy profile (Figure 3.19a). Starting from 2200 ns the free-energy difference between the bound and the unbound states fluctuates around a value of -8.0 kcal/mol with a variation of ±1 kcal/mol. The calculated binding free-energy revealed that the KG-721 ligand has a lower binding affinity than THS-020 (-11.8 kcal/mol), in agreement with the experimental data: ΔG_{exp} (KG-721) = -6.9 ± 0.1 kcal/mol, ΔG_{exp} (THS-020)= -7.9 ± 0.5 kcal/mol. During the simulation, we observed multiple binding and unbinding events and the hill heights decrease toward 0 (Figure 3.19 b and c).

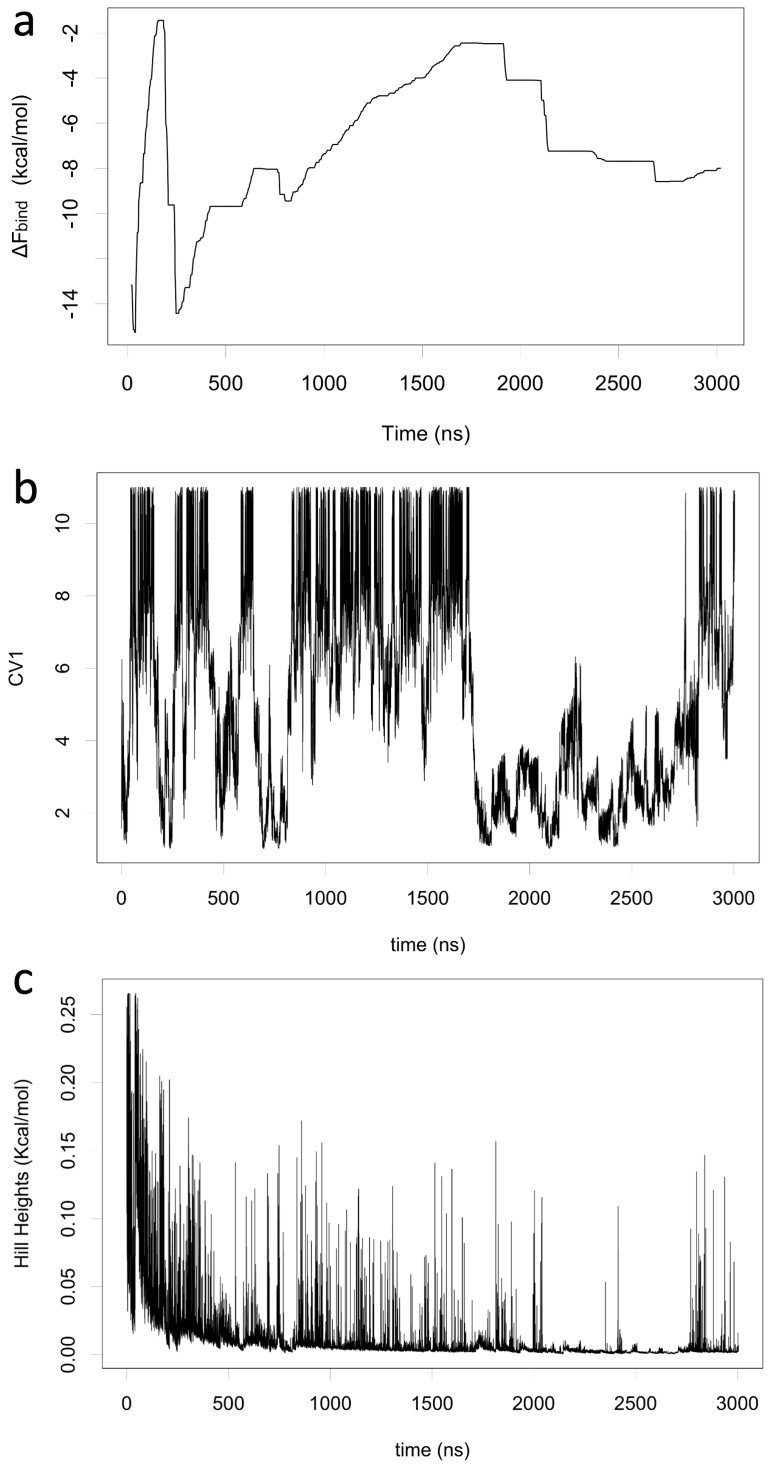

Figure 3.19 Results of KG-721 MetaD simulation. a) one-dimensional projection of the binding free-energy values associated to the path 1; b) instantaneous values of CV1 (s(R)); c) plot of the hill heights during the simulation time.

The final FES obtained for this system is shown in Figure 3.20. Even in this case, we identified several minima, in Table 3.5, and we used the centroid of the most populated cluster in each minimum, as the representative structure of that minimum.

Table 3.5 Coordinates and binding free-energy values of the relevant minima along the CV1 (s(R)) and CV2 (z(R)) for the KG-721 ligand.

Minimum	s(R)	z(R)	ΔF (kcal/mol)
A	1.566 1.699	0.032 0.038	1.15
B	2.421 2.553	0.040 0.046	0.00
C	1.978 2.089	0.115 0.129	2.83
D	1.511 1.699	0.168 0.175	1.18
E	4.097 4.752	0.134 0.153	5.20
F	4.244 4.401	0.015 0.021	4.40
G	6.815 7.008	0.124 0.143	6.15

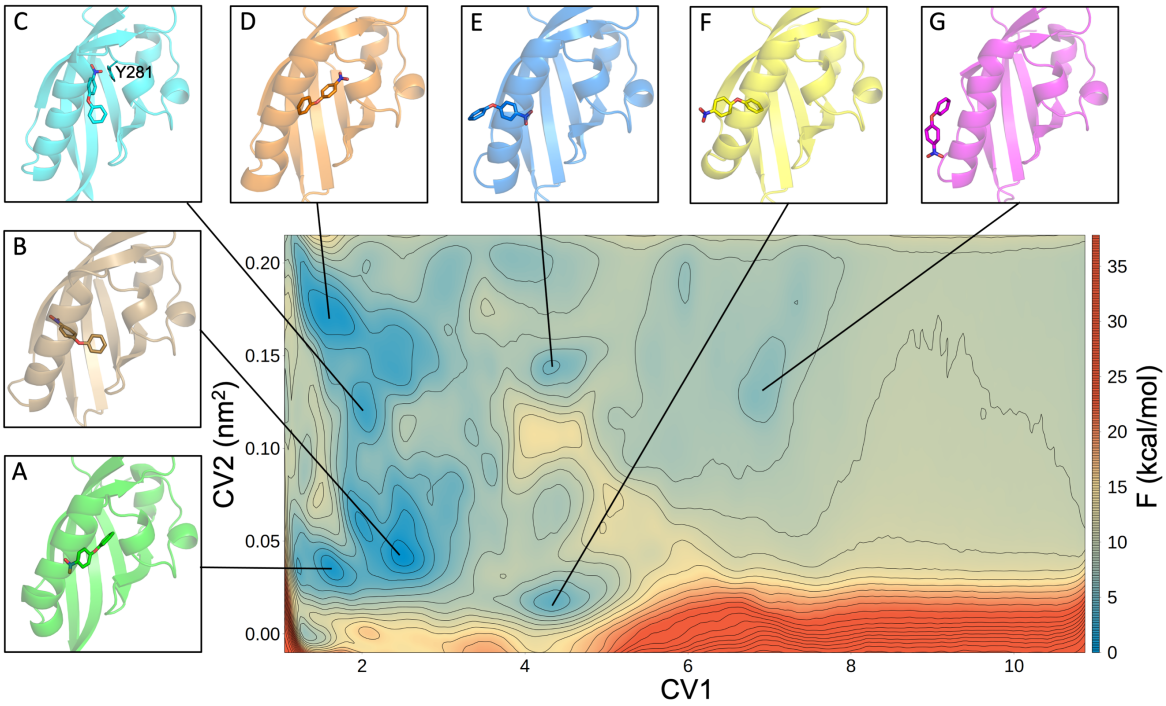

Figure 3.20 Free-energy surface obtained from the PCVs approach for the binding/unbinding of the KG-721 ligand. The isolines are drawn using 1.5 kcal/mol spacing. The 3D structures of the centroids of the main minima are reported with different colors: the protein is represented as cartoons and the ligand as sticks. The black lines indicate the corresponding minima in the FES.

Again, following the CV1, three regions can be distinguished: the bound state, between 1 and 3; the intermediate states, between 3 and 7; and the unbound state, from 7 and on. The region around the bound state displays a multiplicity of alternative binding geometries and does not allow to distinguish a favorite bound minimum. Minima from A to D can be associated to alternative bound states in which the ligand rotates within the binding cavity. In particular, in minimum A the ligand is oriented with NO_2 towards the most polar part of the cavity (S292, S304, and Y307 residues, at the entrance of the cavity) and the phenyl ring towards the apolar part (F244, F254, I261 residues, at the bottom of the cavity), as expected (Figure 19, right panel). Even in minimum B, NO_2 is oriented towards the polar region but the phenyl ring is lightly bent with respect to the other ring. Moving up to higher CV2 values, minima present different orientations of the ligand inside the cavity: minimum C is stabilized by a hydrogen-bond between NO_2 and the Y281 residue; in minimum D, ligand is rotated 180° with respect to minimum A and does not show stable hydrogen-bonds with the protein.

While we previously observed that, in the deepest minimum (A), THS-020 well overlaps the experimental binding geometry (Figure 3.21, left panel), the minimum A of KG-721 (Figure 3.21, right panel) is the most similar to the starting docking pose (which values in the CVs subspace are s(R)=1.3 and z(R)=0). Indeed, the RMSD between the centroid of minimum A and the docked pose is 2.45 Å, indicating that the two conformations are quite different.

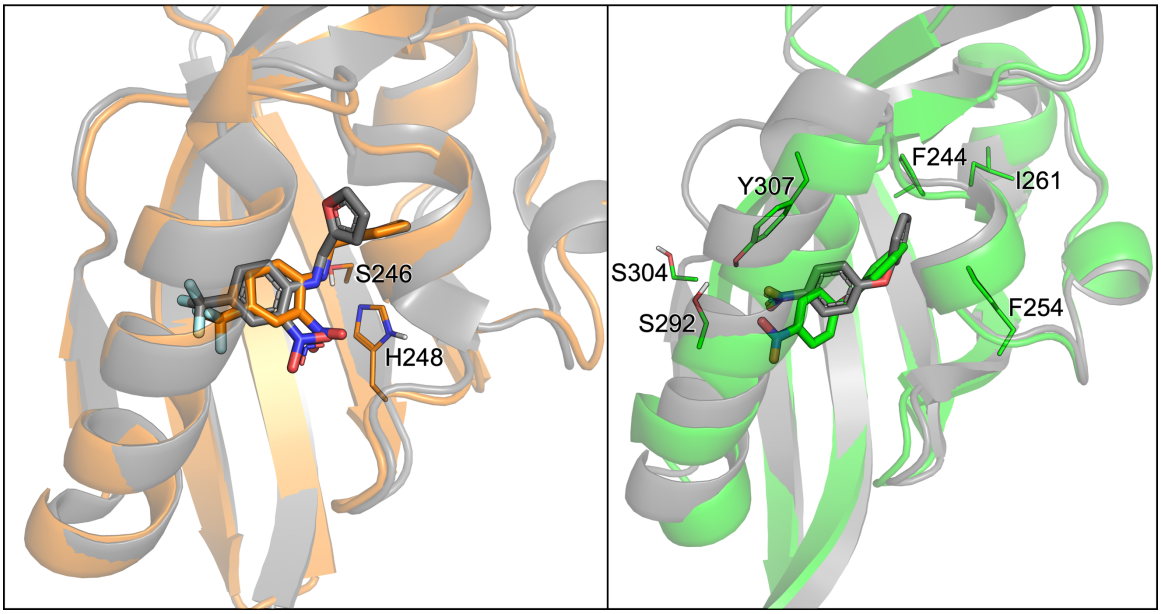

Figure 3.21 Comparison between minimum A and the starting structure for the two ligands. On the left: for the THS-020 ligand, the overlay of minimum A, in orange, with the crystallographic structure of the complex, in grey. On the right: for the KG-721 ligand, the overlay of minimum A, in green, with the docking pose, in grey.

In the case of KG-721, that is not congeneric with any of the co-crystallized HIF-2α ligands[100,101,103], the ensemble-docking strategy was not sufficient for the correct definition of the binding mode. In light of our results, we underline the importance of including protein flexibility more completely. Our results indicate that MetaD calculations are not influenced by the inaccurate starting conformation of the complex but lead the system to evolve to a more stable conformation. Therefore, this technique appears a promising tool in cases where structural information for congeneric ligands is not available.

3.4 Conclusions

Modeling the pathways for ligand binding to the HIF-2α PAS-B domain represents a non-trivial task due to the buried nature of the binding cavity that suggests significant protein conformational changes may occur upon ligand access. The computational protocol here proposed effectively combines two promising methods based on enhanced-sampling MD. Steered MD simulations are used to identify the preferred unbinding pathway among alternative ones and to guide the construction of the reference path for the subsequent step. On the other side, Metadynamics, with the Path Collective Variables formalism, is used to obtain a more rigorous characterization of the free-energy surface and to calculate the binding free-energy value.

By applying this approach to elucidate the binding process of two different ligands of HIF-2α, we obtained the correct binding affinity scale, according to the experimental data available, and we identified minima in the FES that clearly depict the bound state(s) and the intermediate states characteristic of each ligand. Moreover, the method was effective in leading the system to evolve to the most stable binding conformation, starting either from an X-ray structure of the ligand-protein complex or from a docking pose. Therefore, it appears a promising tool also in cases where reference structural information is lacking.

Given the recent discovery of HIF-2α as a pharmaceutical target for cancer therapy, the proposed computational approach based on enhanced-sampling MD appear to be an invaluable tool to investigate the binding process of different ligands, thus contributing to the development of successful drug design projects. The results obtained here also encourage us to extend applications to other binding mechanisms of bHLH-PAS proteins, including significant targets, such as the AhR, for which no experimental structural information on the ligand-bound states is available.

PATHDETECT-SOM: A NEURAL NETWORK APPROACH FOR THE IDENTIFICATION OF PATHWAYS IN LIGAND BINDING SIMULATIONS

4.1 Introduction

As discussed in chapter 1, enhanced sampling methods are now routinely used to simulate the complete binding and/or unbinding events. In particular, PP methods, presented in chapter 2, have the advantage of explicitly simulating key molecular events, such as the protein conformational changes that facilitate ligand access to the binding cavity, and the formation of intermediate states. All the above information is fundamental to suggest appropriate modifications of hit compounds in drug-design studies. However, PP methods generally require an extensive sampling of binding/unbinding events to obtain an accurate description of the energy landscape of the process based on reliable statistics. It follows that many events have to be analyzed through several simulation replicas, or with a single simulation that describes several re-crossing events. The large amount of data from different replicas or events calls for better automated tools to analyze all the simulated events at once and to provide a

clearly interpretable summary picture of the differences in the sampled pathways. We suggest the use of Self-Organizing Maps (SOMs)[148] to handle such complex sets of data. A SOM is a type of artificial neural networks useful for effective identification of patterns in the data[149,150,151] and has been widely used in many fields[152,153]. The most interesting property of a SOM is that it performs a dimensionality reduction by mapping multidimensional data on the SOM grid, retaining topological relationships between neurons, i.e., keeping similar input data close to each other on the map[149]. Several applications of SOMs to the analysis of biomolecular simulations can be found in the literature[154,155,156], ranging from comparison of the dynamics of different mutants[157], clustering of ligand poses in virtual screening[158], binding site identification[159], identification of blocks for structural alphabets[160,161,162] and conformational analysis of loop opening[163]. More recently, we applied SOMs to the reconstruction of protein unfolding pathways on the basis of several sMD simulation replicas[164].

During the PhD project we designed, implemented and tested PathDetect-SOM (Pathways detection on SOM), a SOM-based protocol for the analysis of ligand binding/unbinding pathways derived from MD simulations with PP methods. Taking advantage of the properties of SOMs, the tool is able to generate a model that clearly highlights differences in the pathways sampled along a simulation or in different replicas. The protocol makes it possible to obtain a synthetic view of the sampled conformational space by highlighting the relevant states, to trace the pathways followed by the system on the SOM, and to derive a network model that provides a meaningful representation of the binding/unbinding pathways.

We applied this protocol to three study cases selected to represent PP simulations with different characteristics. The three study-cases are briefly presented in the following:

The first case regards the unbinding of the THS-020 ligand from the HIF-2α PAS-B domain, studied through sMD simulations as reported in chapter 3. The simultaneous evolution of the replicas (due to the constant velocity of the bias) and the use of a

directional Collective Variable (CV) make this study-case simple and optimal for the initial testing of some parameters of the tool (tests are discussed in paragraph 4.3).

The second study-case refers to the unbinding of the GC7 ligand from the Deoxyhypusine synthase (DHS). DHS is an enzyme responsible for the post-translational hypusination of the eukariotic initiation factor 5A (eIF5A) that controls cell proliferation and it has been linked to cancer[165]. The involvement in pathogenesis together with the high specificity and functional relevance of the hypusination reaction have made this system an important and promising therapeutic target, stimulating the design and development of inhibitors able to target the hypusination process, including the GC7 ligand. In particular, GC7 interacts in a specific binding pocket of the DHS and completely blocks its activity; however, its therapeutic use is limited by poor selectivity and restricted bioavailability. In a recent work[166], a comparative study has been performed between the unbinding pathways in the human DHS (hDHS) and in the archaeal DHS from crenarchaeon Sulfolobus solfataricus (aDHS), by using an approach inspired to Infrequent MetaD. As in the previous study case, several replicas are performed but, differently from sMD, the system evolves along the selected CV with a series of small forth and back movements that fill the free-energy basin. As a result, there is no correspondence between the simulation times of different replicas. Moreover, the type of CV chosen in this case is non-directional and may provide very different unbinding paths.

The third study-case still concerns the THS-020-HIF-2α complex, but the binding/unbinding of the ligand is studied through a single long MetaD simulation, in which several binding and unbinding events are sampled (as already reported in chapter 3). In this case, the simulation evolves in all the directions according to two selected CVs and the ligand has greater freedom than in the previous cases.

For all the processes, the results provided not only a simple schematic representation of the ligand binding/unbinding pathways, but also hints about the thermodynamics and/or kinetics of the process.

4.2 Methods

Overview of the protocol

PathDetect-SOM is a modular command-line tool based on a three-step protocol (see Figure 4.1):

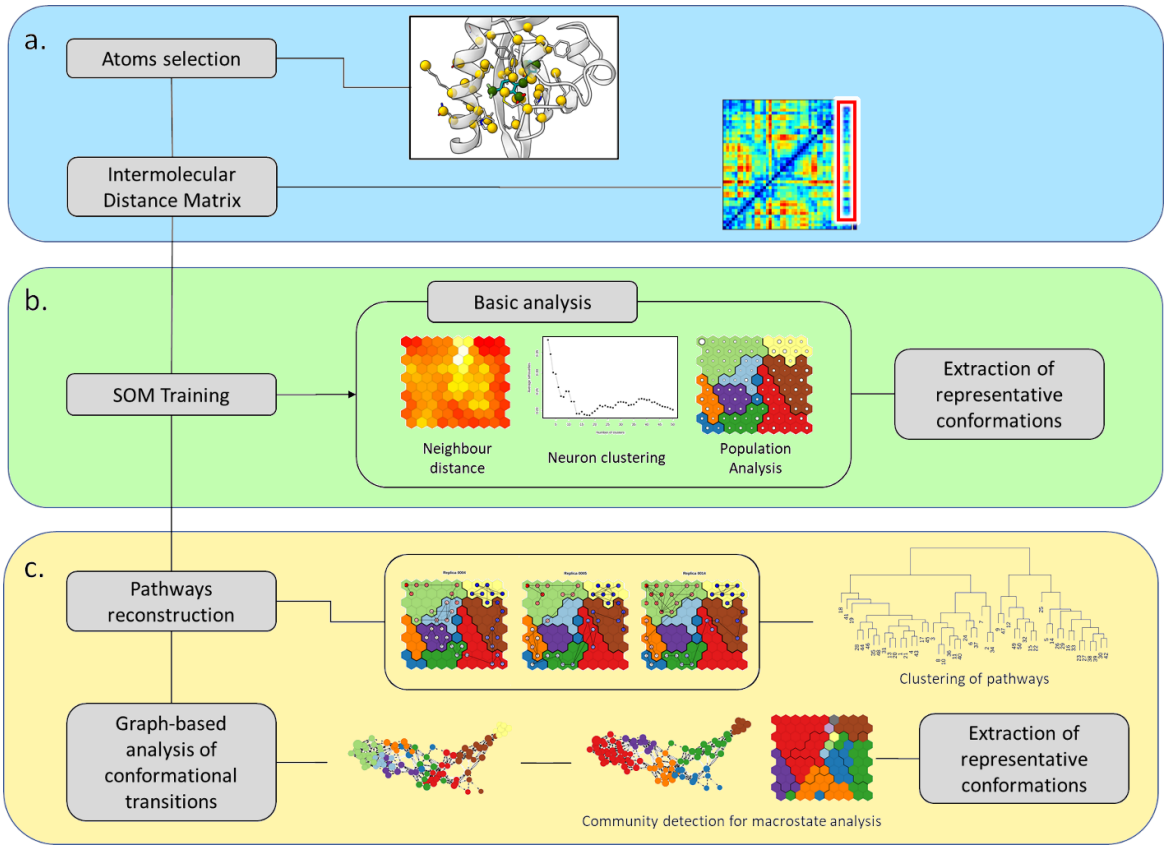

Figure 4.1 Flowchart of the PathDetect-SOM protocol for ligand binding studies. Data preparation (a); map training and analysis (b) and pathways analysis (c).

a) The user selects a set of features best describing ligand conformations along the process. If a set of protein and ligand atoms is provided, the tool will automatically compute the intermolecular distances.

b) SOM is initialized and trained with the input vectors containing the values of the selected features for all the simulation frames. Each frame is considered as a data point and assigned to the neuron with most similar feature values. During the training process the feature values of a neuron and its neighbors are adjusted toward the values in the input vector assigned to that neuron. The final prototype vector of each output neuron summarizes

the conformations associated to the neuron and groups of similar conformations are mapped to neighboring neurons. In addition, to offer a more concise picture of the map, after training the neurons are also grouped to a relatively small number of clusters and the representative conformation of each cluster is saved. Population analysis and average properties can then be visualized on the trained SOM.

c) The pathways followed during the simulation can be directly traced on the SOM, reconstructing the binding/unbinding pathway. This representation facilitates the identification of regions of the map exclusively sampled by specific simulations. In turn pathways can be clustered to recover dominant binding events. Finally, a graph-based representation of transitions can be built from the transition matrix calculated at the neuron level. Community detection on this graph can highlight putative macrostates.

PathDetect-SOM is distributed as an R script available under GNU General Public License at https://github.com/MottaStefano/PathDetect-SOM. The repository includes a brief guide and tutorial material based on sample trajectories from the first study case.

Data preparation

The feature selection is a key step for SOM training. Several features can be used to train the SOM (for example the simple cartesian coordinates of a set of atoms, more details are discussed in paragraph 4.3). However, the intermolecular distances are the most suitable choice to accurately describe the ligand-receptor reciprocal orientation. A set of receptor and ligand atoms is chosen for the computation of intermolecular distances. Selected atoms should describe both the binding site and the mouth at the entrance of the binding site. Ideally, both atoms from backbone and from large or polar/charged sidechains should be included when the side chain dynamics and interactions are relevant for binding. Similarly, selected ligand atoms should well describe the core molecular structure and all the relevant lateral groups. The user can provide the filtered trajectory with the coordinates of the chosen atoms in the form of an xvg file, easily obtained using the GROMACS gmx traj command. A capping value

is applied to the distances to avoid that training is dominated by information on the unbound states (more details are discussed in paragraph 4.3). Details on the atom selection and capping values for the study-cases here presented are summarized in Table 4.1.

Table 4.1 : Details for feature calculations of each system

System	Atom selection	Capping value (nm)
HIF-2α – THS-020	S246 - OG; H248 - NE2; H248 - CA; M252 - CE; M252 - CA; F254 - CZ; F254 - CA; A277 - CB; F280 - CA; Y281 - OH; Y281 - CA; N288 - CG; N288 - CA; M289 - CE; M289 - CA; K291 - NZ; K291 - CA; S292 - OG; H293 - NE2; H293 - CA; N295 - CG; N295 - CA; L296 - CG; L296 - CA; V302 - CB; V303 - CA; S304 - OG; G305 - CA; Q306 - CA; Y307 - OH; Y307 - CA; M309 - CE; M309 - CA; T321 - OG1; T321 - CA; I337 - CB; C339 - SG; C339 - CA; N341 - CG; N341 - CA	1.2
System	**Atom selection**	**Capping value (nm)**
DHS – GC7	Chain A: K260 - CA; H261 - CA; H261 - NE2; N265 - CA; L268 - CA; L268 - CG; M269 - CA; E284 - CA; E284 - CD; G287 - CA; S288 - CA; D289 - CA; D289 - CG; A292 - CA; E296 - CA; E296 - CD; W300 - CH2; K302 - NZ Chain B: N79 - CA; N79 - CG; G106 - CA; E109 - CD; E110 - CD; N137 - CA; R138 - CA; I139 - CA; G140 - CA; Y149 - CA; D211 - CA; S213 - CA; S213 - OG; D216 - CA; D216 - CG	1.6

Map training

The selected features are used to train the SOM using an iterative approach. The map is initialized by assigning random values of the feature vectors to each neuron. In each training cycle the input vectors representing the single conformations are presented in random order to the map and assigned to the neuron with closer feature values, also called best matching unit (BMU). The feature values of the BMU and its neighbors are modified to be closer to the values of the input vector. The magnitude

of the modification decreases with the distance from the BMU and along the training. At the end of the iterative process the resulting SOM preserves topological relationship between neurons, keeping similar original input data close on the map. In a second step, the neurons are further grouped in a small, but representative, number of clusters by agglomerative hierarchical clustering using Euclidean distances and complete linkage. For each system, the optimal number of clusters can be selected on the basis of silhouette profiles as show inFigure 4.2.

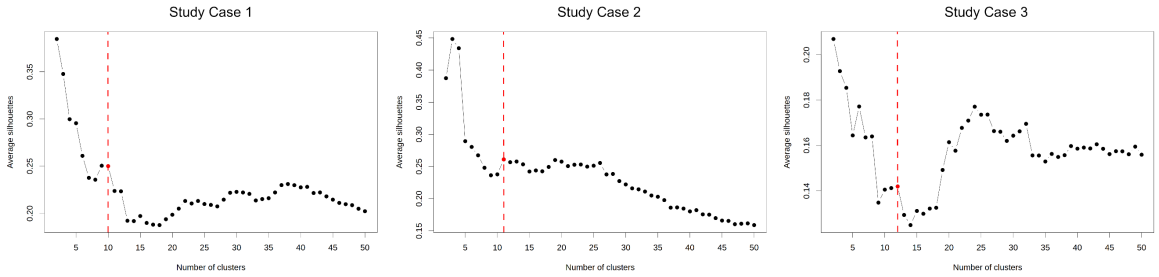

Figure 4.2 Silhouette profiles for the three study cases. The optimal number of clusters (red) was chosen as the one with the highest silhouette score in the range 9-15.

A representative structure for each neuron is saved; this is defined as the structure with the feature vector closest to the neuron vector. For each cluster, a representative neuron is also chosen as the one with the feature values closest to the weighted-average feature vector of the neurons belonging to that cluster. In the latter case, the average was performed using the population of each neuron as weight.

In the present work 10x10 sheet-shaped SOMs with hexagonal lattice shape and without periodic boundary conditions were trained over 5000 training cycles. The neurons were further grouped in a small, but representative, number of clusters, different for each study-case, using the cluster analysis approach outlined above.

Path analysis

The trained SOM captures the conformational space of several trajectories in a topological map. Therefore, it is possible to reconstruct the path explored by each simulation on the map. Pathways are traced on the SOM based on the annotation of the BMU associated to each frame of the simulation. The resulting SOM pathways were also clustered by agglomerative hierarchical clustering using average linkage. Two different distance metrics are implemented in the PathDetect-SOM tool: a time-

dependent and a time-independent distance. In the time-dependent version the distance between the SOM pathway of two simulations is defined as the average distance of the BMUs of each couple of frames. The distance between two BMUs is defined as the Euclidean distance between the position of the neurons on the map. This distance was used also in a previous work by the authors[164]. In the time-independent version, for each frame of the simulation, the minimum distance between the BMU of the first and the second simulation is computed and averaged over the number of frames. This approach provides a framework to compare simulations evolving at different speeds such as those presented in study-case 2. For this type of simulation, indeed, frames to be compared are not at the same position along the replicas, due to the different evolution of the simulations. Comparing each frame with the closest frame of the second replica, is a time-independent way of performing a distance calculation between two pathways.

An approximate transition matrix between each pair of neurons can be computed from the time-dependent distance approach. The matrix is then transformed into a row stochastic matrix and a graph is built with nodes representing the neurons and edges with weight proportional to the negative logarithm of the transition probability between the corresponding neurons. Communities of nodes can be detected and in the present work we used the walktrap algorithm[167], but other methods can easily be applied. A neuron representative of each community is selected as the one with the highest eigenvector centrality score in the subgraph which only contains nodes belonging to the community.

In this work, for the third study-case, a commitor analysis was performed using the R library markovchain[168,169]. This analysis computes the probability of hitting a set of states A before the set B starting from different initial states. In this case the two extremes were the bound and unbound states. All the analyses were performed in the R statistical environment using the kohonen package[170,171] for the SOM training and igraph package[172] for graph construction and analysis.

The main results of the work reported is this Chapter are reported in ref[173].

4.3 Results

The PathDetect-SOM protocol, developed for the analysis of ligand binding/unbinding pathways, is implemented into a command-line tool with the capability to build a SOM representation of the conformations sampled during the MD simulations. Taking advantage of the SOM topological ordering, the tool offers the possibility to visually represent pathways sampled during different events/replicas in a clear 2D representation. Finally, the geometric microstates identified by the SOM (neurons) can be represented as a graph model, built from their transition probabilities. The graph provides a clear representation of the pathways followed during the simulations, facilitating the identification of alternative routes. Community detection on the graph generates a state model analogous to kinetic partitioning.

In the following sections, we present details regarding the selection of parameters for the SOM training, and the application of the protocol to the three study-cases introduced in paragraph 4.1.

Selection of optimal parameter values for SOM training

The application of the PathDetect-SOM tool requires the choice of a series of parameters for the initial training of the SOM. Therefore, some tests were performed to select the optimal parameters, based on preliminary analyses of the sMD simulations of THS-020 unbinding from HIF-2α (the first study-case). Here, details concerning the choice of features describing the conformations and the associated distance measures, the choice of a capping value for the distances, and the type of periodic boundary conditions are discussed.

Features: The PathDetect-SOM tool can train the map according to different features representing the input conformations from MD trajectories. As each feature has a specific relevant distance measure, user options are based on the type of distance measure: the RMSD (Root Mean Square Deviation) or the dRMSD (distance RMSD). In the case of the RMSD, the SOM is directly trained with atomic coordinates (in the case under study, the coordinates of the ligand heavy atoms). This means that simulation frames should be pre-aligned (in this case, on protein Cα atoms). In the

case of the dRMSD, the SOM is trained with a set of distances (in this case, the intermolecular distances). Results from each measure are reported in Figure 4.3.

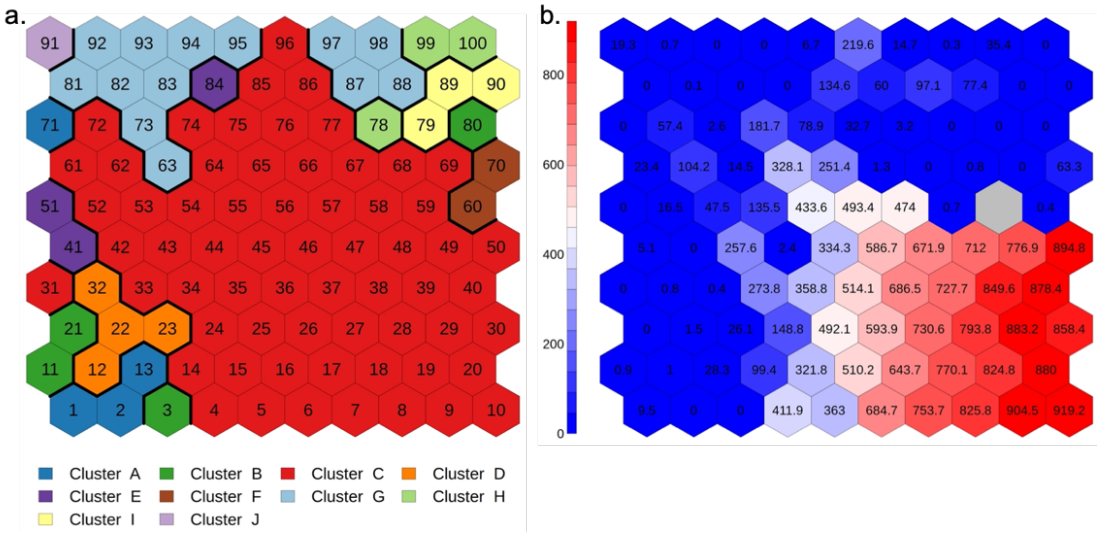

Figure 4.3 SOM trained using RMSD. (a) clustering of neurons; (b) protein-ligand contacts plotted on SOM. Each neuron is colored according to the average number of contacts in the frames belonging to it. High number of contacts are depicted in red and assigned to neurons describing the bound state.

The choice of the RMSD (Figure 4.3a) generates a large cluster (C) whose neurons describe both bound and pre-bound states. This becomes evident by mapping the protein-ligand contacts on the SOM (contacts were considered when two atoms are closer than 4.0 Å, Figure 4.3b) and observing that cluster C includes neurons that exhibit both high and low numbers of contacts. When the ligand is outside the cavity (low number of contacts, blue neurons in Figure 4.3b), the variability of its conformations is so wide that it hides the changes in the ligand-protein distances that occur along the binding pathways. The results shown in the main text were obtained using dRMSD as distance type and generated a more consistent and informative description of the pathways. When studying a protein-ligand binding process, the intermolecular distances (dRMSD) are the best choice, because the SOM is directly trained on the information relevant for the process as recorded by the changes in intermolecular interactions and, in addition, there is no requirement for preliminary structural superposition.

Capping value: This parameter may be used to assign a given fixed value to all the distances between the selected ligand and protein atoms that are greater than a user-specified value. SOMs were trained with different capping values: no capping, 8 Å and

12 Å. To understand how this parameter affects the assignment of simulation frames to the neurons, and consequently the pathway description, the number of ligand-protein contacts was calculated for every frame of the simulation (contacts were considered when two atoms are closer than 4.0 Å, Figure 4.4).

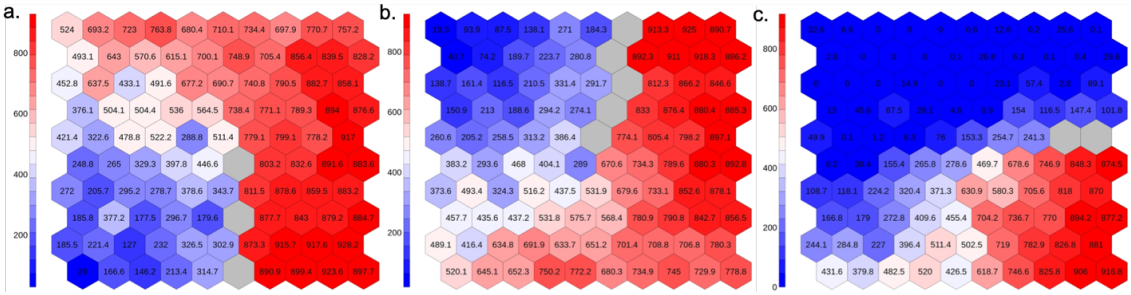

Figure 4.4 SOMs trained with different capping values for the distances. Number of protein-ligand contacts are plotted on the SOMs: each neuron is colored according to the average number of contacts in the frames belonging to it; high number of contacts are depicted in red and assigned to neurons describing the bound state. (a) SOM trained with 8 Å as capping value, (b) SOM trained with 12 Å as capping value, (c) SOM trained without capping value.

Using the lowest capping value, 8 Å (Figure 4.4a), the neurons describing the bound state cover most of the map, and only few neurons describe the last portion of the unbinding process. Indeed, a jump in the number of contacts between the neuron containing the unbound state (bottom left corner of the map) and its neighbors is visible. Without the use of capping (Figure 4.4c), the neurons describing the unbound state cover more than 50% of the map, with poor description of the recognition process. On the contrary, using a value of 12 Å (Figure 4.4b), a balanced description of all steps of the binding/unbinding process is observed, and the number of contacts gradually changes across the neurons. Given that the map is sensitive to the number of distances reaching the capping value when the ligand gets in the unbound state, it is advisable to adjust the capping value based on the length of the binding pathway within the cavity. A good starting value for the capping could be the distance between residues lying at the bottom and at the mouth of the cavity.

Periodic boundary conditions: Another parameter that can be chosen by the user is the periodicity of the SOM. If the SOM grid is periodic across the boundaries, the neurons at the right boundary will be neighbors of neurons at the left boundary, as well as those on the top and bottom boundaries. The SOM was trained with and without periodic boundary conditions. The main difference is evident when pathways

are traced on the SOM. The binding pathways pass through the boundary when training is done with periodic boundary conditions (Figure 4.5) leading to difficulties in interpreting the time evolution of the process.

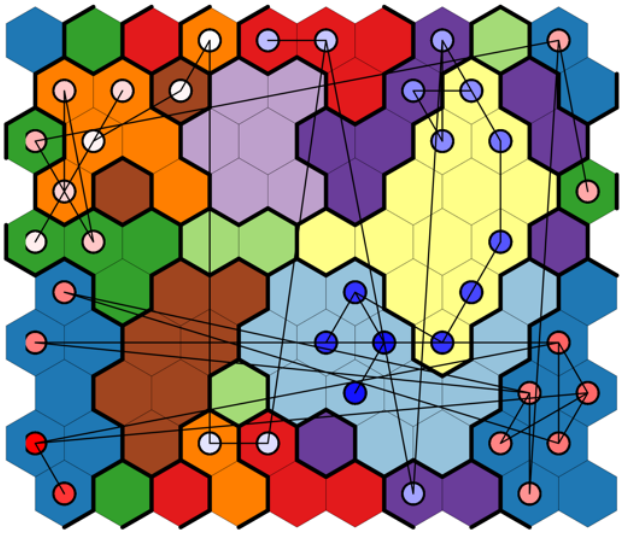

Figure 4.5 An example of pathways traced on a SOM trained using periodic boundary condition.

In addition, the neuron clusters become fragmented across the map, making it difficult to interpret the distribution of the different conformational states in the clusters. The results shown in the main text were obtained without periodic boundary conditions; this choice led to consistent pathways more clearly traceable on the SOM and to easier identification of the different states across the neuron clustering.

Ligand unbinding through multiple replicas with constant velocity pulling

As discussed in chapter 3, we investigated the unbinding of the THS-020 ligand from the HIF-2α PAS-B domain through sMD simulations. 50 constant velocity sMD replicas of 25 ns each were used to pull the ligand along the selected CV, namely, the distance between the center of mass of the aminoacid atoms lining the cavity and the center of mass of the ligand. The simulations analyzed in this chapter are those along the preferred pathway (path 1) identified in the previous work[108] (discussed in chapter 3). All replicas evolved simultaneously, due to the constant velocity of the bias, and along a directional CV. The trained SOM (Figure 4.6 and details in paragraph 4.2) shows a distribution of states that ranges from the initial bound state (top right of the map) to the unbound state (top left).

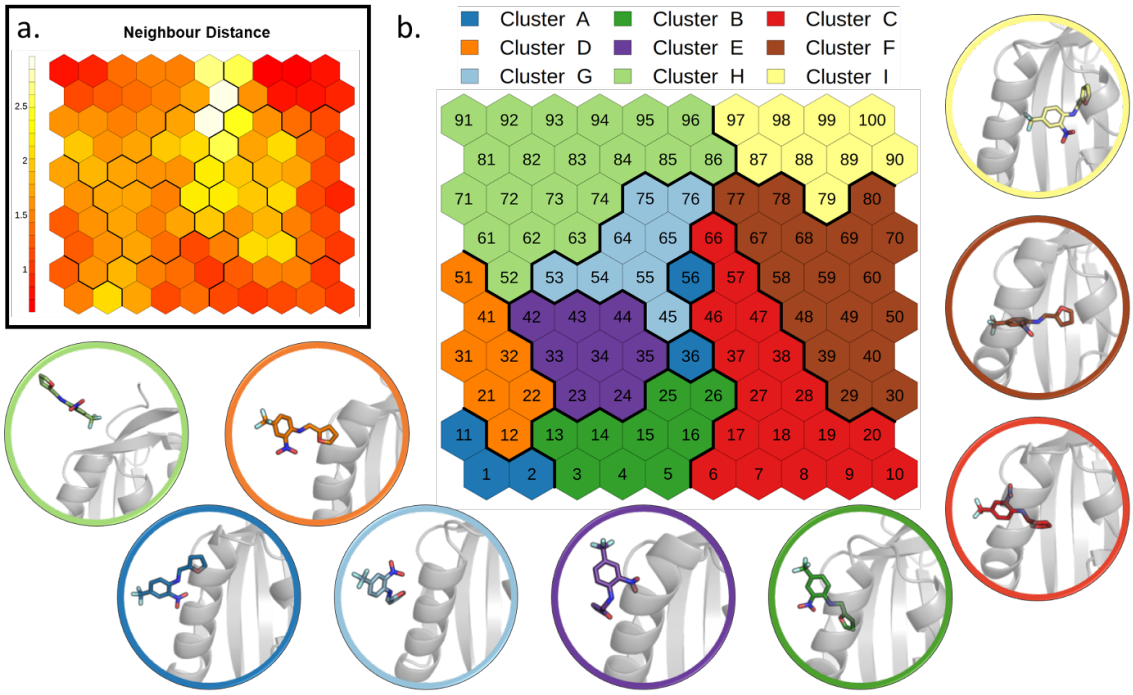

Figure 4.6 SOM analysis of sMD simulations of THS-020 unbinding from HIF-2a: (a) Neighbor distance plot. (b) Clustering of the neurons. The representative conformation of each cluster is depicted in cartoons with ligand in sticks.

The neighbor distance plot (Figure 4.6a) represents the average similarity of a neuron with its neighbors. This map shows a compact group of neurons in correspondence of the bound state and along the right and bottom border of the map. On the contrary, neurons lying at the center of the map displays more heterogeneity. In the cluster analysis of SOM neurons (see paragraph 4.2) we identified 9 clusters that represent the binding geometries explored by the system following the distance CV used for the sMD simulations (Figure 4.6b). The representative conformations extracted from the different clusters help to visualize the relevant states sampled.

The pathways followed by each replica were then mapped on the SOM (Figure 4.7).

Figure 4.7 Tracing of the pathways on the trained SOM for the sMD replicas of THS-020 unbinding from HIF-2a.

They are quite consistent, since they roughly evolve through the same sequence of clusters, in agreement with the high directionality imposed by the method. However, some recurrent unbinding pathways can be identified with slight differences from each other, as also emerges from the dendrogram in Figure 4.8.

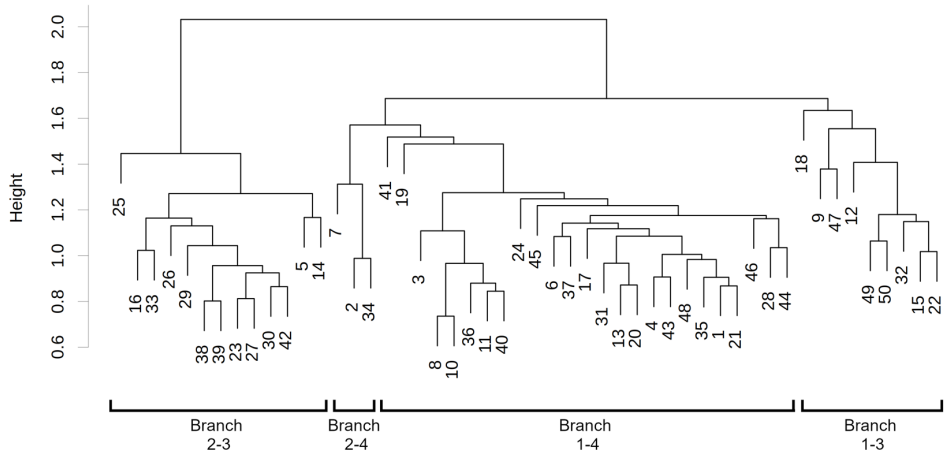

Figure 4.8 Dendrogram of hierarchical clustering of the pathways followed by different replicas for the study case 1 (HIF-2a sMD simulations).

An overview of these pathways is provided by the network graph derived from the transition matrix (see paragraph 4.2), reported in Figure 4.9Figure 4.9a.

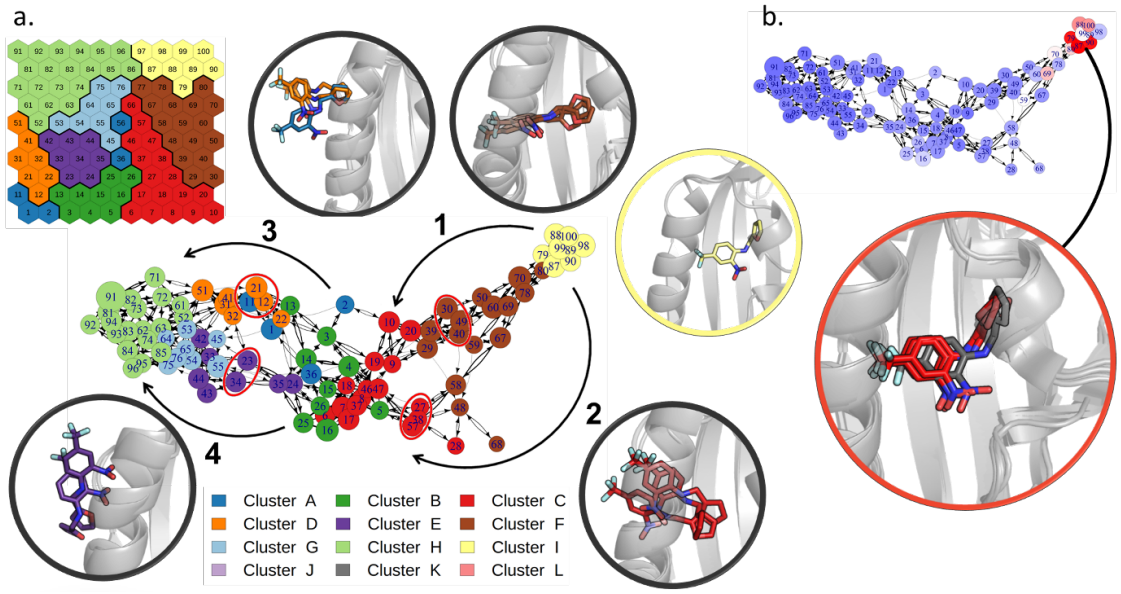

Figure 4.9 Transition network for the sMD simulations of THS-020 unbinding from HIF-2a. a) Transition network with its main ramifications explicitly indicated by black arrows (nodes are colored according to the SOM clusters). The representative conformations of neurons that characterize each branch (red circles in the network) and of the bound state (in yellow), are depicted in cartoons with ligand in sticks. b) Network colored according to the average sMD force of its frames (from blue to red, increasing values of this property), and the representative conformations of the neurons with the maximum forces, superimposed to the bound state (in grey).

All the simulations start from the bound state (top right), in which the ligand presents the nitrobenzene ring parallel to the main helix, with the nitro group pointing toward the lower-side of the cavity. Then some replicas evolve through neurons at the bottom-right of the map (branch 1 of the graph), while others follow pathways closer to the center of the map (branch 2 of the graph). While simulations following branch

1, that was sampled in most of the replicas (34 out of 50), show the ligand slightly rotated along its principal axis, those along branch 2 maintain the ligand in an orientation similar to the bound state, and rigidly translate it along the pathway. When the nitro group reaches the solvent, however, the two branches merge, before a second ramification in the graph appears (branches 3 and 4). Replicas in branch 3 describes a rigid transition of the ligand that maintains the initial bound orientation, while those in branch 4 sample conformations with the ligand rotated and bound to the mouth of the cavity. The two final branches appear equally probable (22 replicas though branch 3 and 28 through branch 4).

Finally, we colored neurons according to the average sMD pulling forces applied to the frames belonging to that neuron (Figure 4.9b). Results show that the pulling of the ligand out of its initial bound state requires the maximum of the force, while the remaining part of the pathway requires less force. We interpreted the peaks of maximum forces as the approximate location of the highest energy barrier to be crossed during unbinding, which corresponds to the energy necessary to pull out the ligand from its initial state.

Ligand unbinding through multiple replicas with a bidirectional sampling

As anticipated in paragraph 4.1, the unbinding pathways of the of GC7 ligand from the Deoxyhypusine synthase (DHS) was previously simulated by using an approach inspired to Infrequent MetaD[55]. The number of contacts between the ligand and the protein binding site atoms was used as a single CV in 30 replicas of infrequent MetaD that were stopped when the ligand reached an unbound state.

By applying the PathDetect-SOM approach to the above simulations, we obtained the trained SOM shown inFigure 4.10. The neighbor distance plot (Figure 4.10a) displays a very compact region on the left side, corresponding to different bound states. All these neurons were grouped together in the neuron clustering phase (cluster A), while the diverse unbound conformations are segregated to the opposite side (Figure 4.10b).

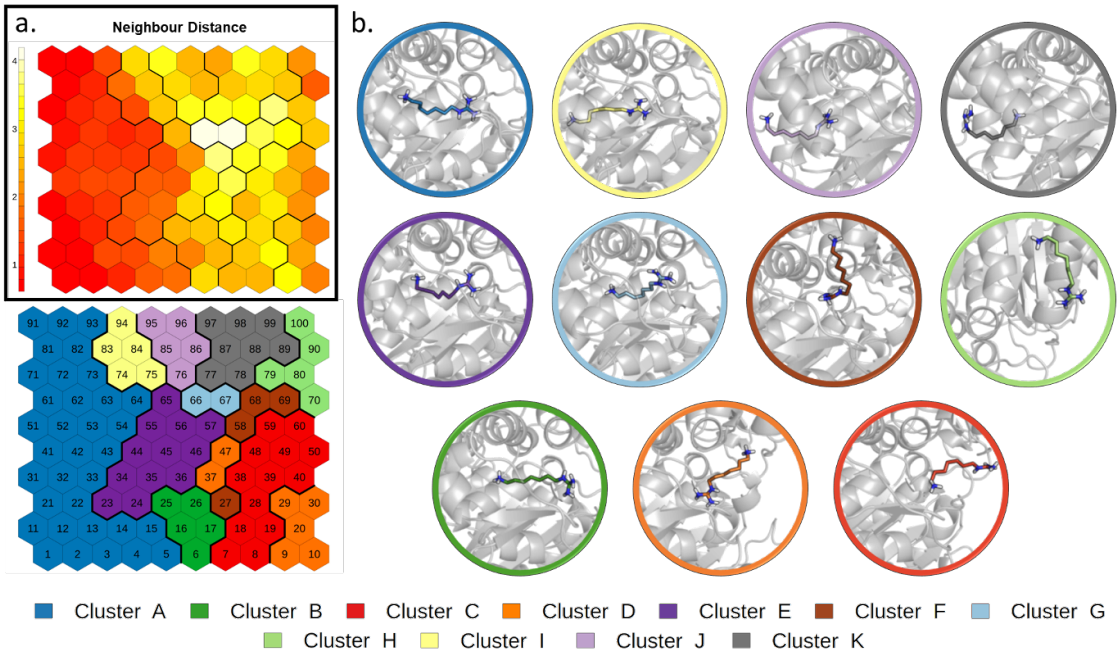

Figure 4.10 SOM analysis of simulations of GC7 unbinding from DHS. (a) Neighbor distance plot. (b) Clustering of the neurons. The representative conformation of each cluster is depicted in cartoons with ligand in sticks.

Due to the nature of these MetaD simulations, where the system evolves along the CV with a series of small forth and back movements, the direct tracing of the pathways on the map may result a little bit confusing (Figure 4.11).

Figure 4.11 Tracing of the pathways on the trained SOM for the MetaD replicas of GC7 unbinding from DHS.

Moreover, given the lack of correspondence between simulation times of different replicas, we needed to perform a time-independent clustering of pathways (see paragraph 4.2), that allows to compare replicas of different length (dendrogram in Figure 4.12).

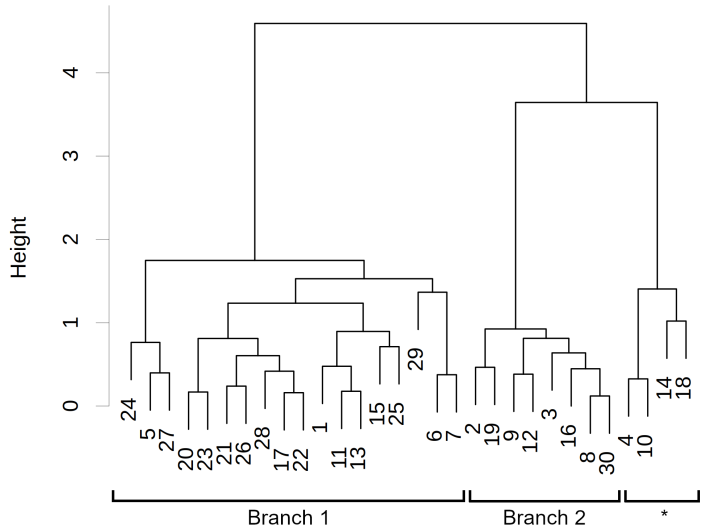

Figure 4.12 Dendrogram of hierarchical clustering of the pathways followed by different replicas for the study case 2 (DHS MetaD simulations). Replicas can be assigned to branch 1 or branch 2 of the network, in good agreement with the clustering, except for four replicas indicated with (*) in the dendrograms. Most of these replicas did not reach a completely unbound state.

Two distinct types of pathways arise from this analysis. Building a network from the transition matrix, as in the previous study-case, made the differences between the two pathways more evident (Figure 4.13a).

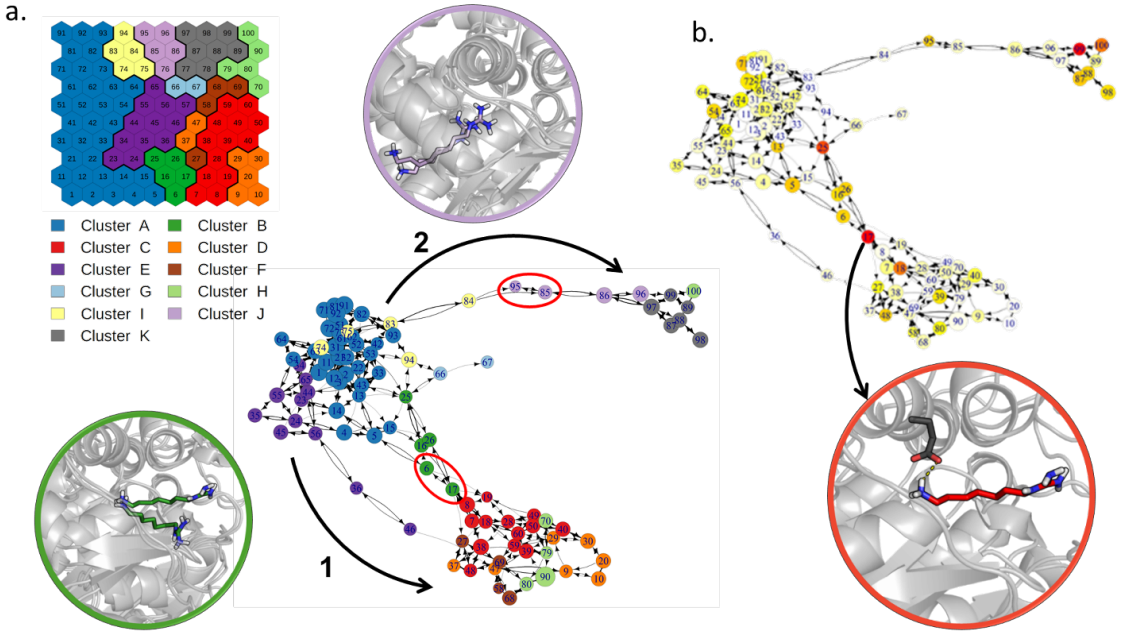

Figure 4.13 Transition network for the simulations of GC7 unbinding from DHS. a) Transition network with its ramification explicitly indicated by black arrows (nodes are colored according to the SOM clusters). The representative conformations of neurons that characterize each branch (red circles in the network) are depicted in cartoons with ligand in sticks. b) Network colored according to the node betweenness centrality (from white to red, increasing values of this property), and the representative conformations of neuron 17, bottleneck for pathway 1.

The two pathways (branch 1 and 2 of the network, in Figure 4.13a), lead to different neurons, all describing unbound states. The separation of the unbound states in different neurons is due to the ligand exiting from the two opposite sides of the binding site. Compared to the previous case, this graph is more densely connected due to the bidirectionality of the sampling during the MetaD simulation. Most of the simulations (70%) evolve through branch 1 (Pathway A in the original work[166]) in which the ligand escapes from the side of its guanidine group. The remaining replicas (30%) proceed through an opposite pathway, indicated as branch 2 in the graph, in which the ligand exits from the side of its ammino-group (Pathway B in the original work[166]). Interestingly, most of the simulations following branch 1 pass through neuron 17, a node with a high value of betweenness centrality (Figure 4.13b). As betweenness is calculated as the number of shortest paths through a node[174], neuron 17 is a critical conformation to observe the bound/unbound transition. The representative conformation of this neuron shows the characteristic of the intermediate state hypothesized in the original work[174], namely, a stable salt bridge of the ligand primary ammine group with Glu137.

Ligand binding/unbinding through a single metadynamic simulation

As a third study-case, we applied the PathDetect-SOM protocol to a single MetaD simulation of ligand binding. The system under investigation is the same as in the first study-case: the THS-020 binding to HIF-2α. As presented in chapter 3, starting from the sMD simulations, we built a path CV and used well-tempered MetaD to enhance the sampling along the selected CV and to reconstruct the free-energy landscape of the process[108]. During the 1.8 μs of MetaD simulation, we observed a high number of binding and unbinding events (Figure 3.14A).

The trained SOM (Figure 4.14 and details in paragraph 4.2) presents the starting bound conformation in the top-left corner (cluster L), and the completely unbound conformation in the top-right corner (cluster G).

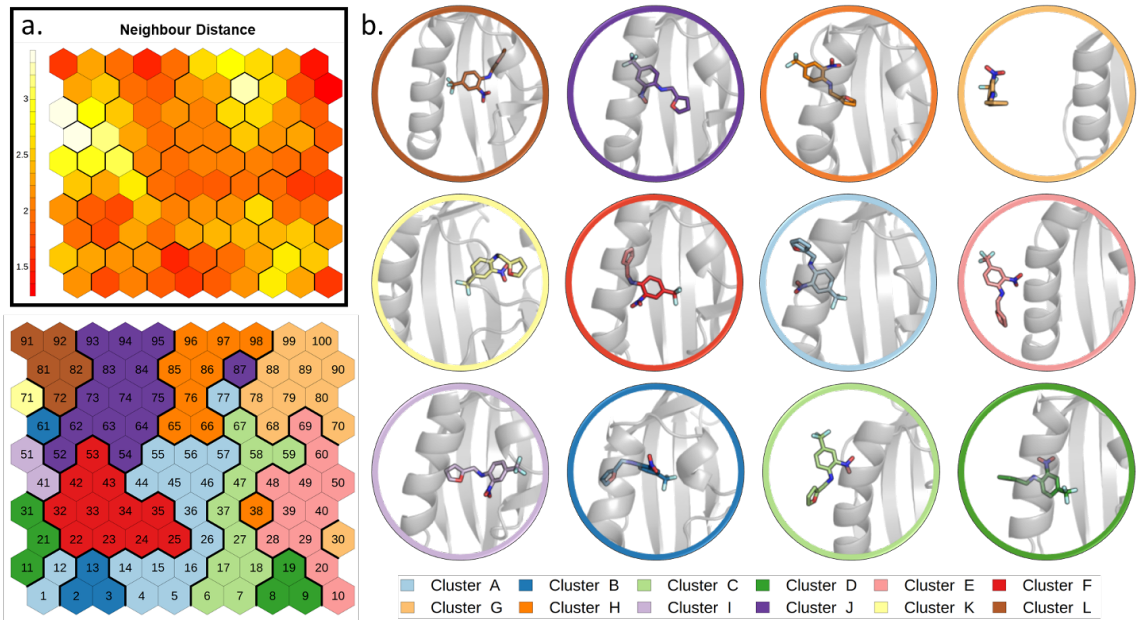

Figure 4.14 SOM trained with MetaD simulations of THS-020 binding to HIF-2a. (a) Neighbor distance plot. (b) Clustering of the neuron vectors. The representative conformation of each cluster is depicted in cartoons with ligand in sticks.

Due to the conformational freedom along the z(r) CV of the path CV (that represent the distance from the reference path), the ligand can also rotate and sample alternative bound conformations. This is the case of cluster I, that contains conformations in which the ligand is rotated of 180° with respect of the X-ray starting structure.

For the sake of comparison with the free-energy landscape previously identified by the MetaD calculation[108] (discussed in chapter 3, Figure 3.16),we mapped the frames belonging to each free-energy basin on the SOM (Figure 4.15).

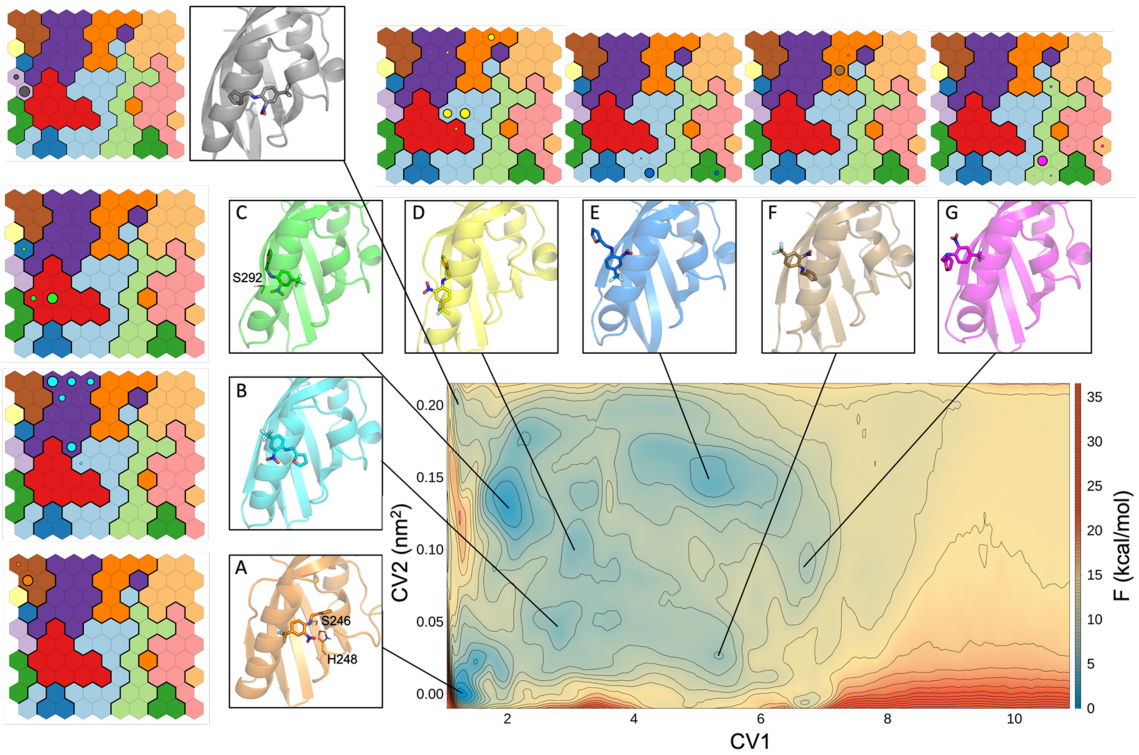

Figure 4.15 Mapping of frames belonging to each free-energy basin on the SOM for the study-case 3. Letters in squares correspond to the free energy states identified in the original work[108] (see Figure 3.16). Cluster I (top left, grey) do not correspond to any of the original lowest free-energy states of the map.

We found that conformations belonging to each of these basins generally map in few close neurons, belonging to a same cluster on the map.

In this study-case, the direct tracing of pathways on the SOM is difficult due to the unique long simulation that samples multiple binding/unbinding events

However, the transition network analysis proposed in the PathDetect-SOM protocol is capable of providing a clear representation of the pathways sampled during the MetaD simulation (Figure 4.16a).

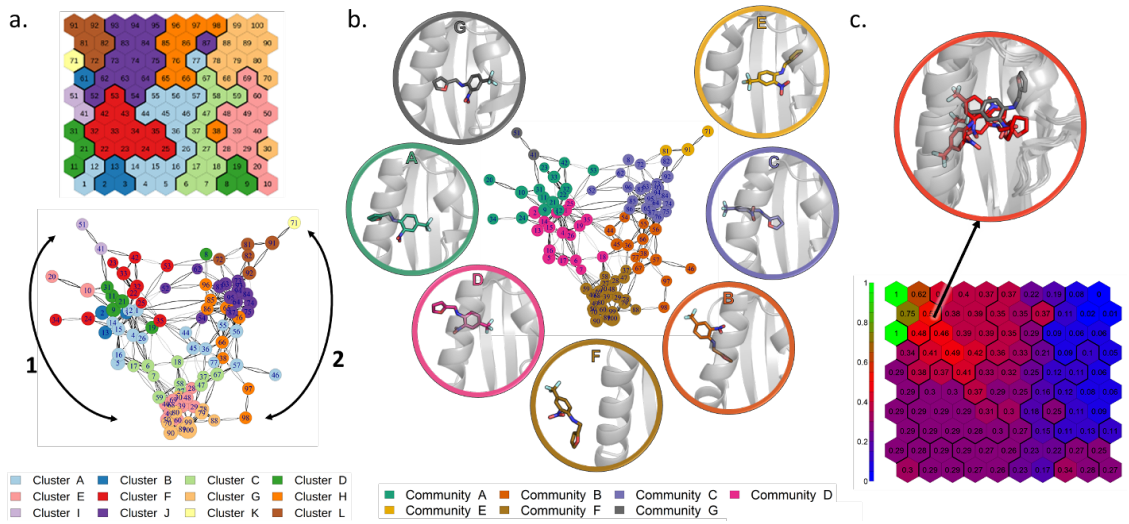

Figure 4.16 Transition network for the MetaD simulation of THS-020 binding to HIF-2a. a) Transition network with main pathways indicated by black arrows (nodes are colored according to the SOM clustering). b) Communities identified by the walktrap method represented on the network (nodes are colored according to the different communities). The representative conformations of the communities are depicted in cartoons with ligand in sticks. c) Committor probability analysis. The representative conformations of neurons with committor probability of about 0.5 are reported in red stick and X-ray starting conformation in grey sticks.

As shown in Figure 4.16a, there are two main branches: branch 1 connects the crystallographic-like bound conformation to the unbound state, while branch 2 follows the unbinding of an alternative binding mode (cluster I). Only a small number of connections between the two branches is present, indicating that the ligand cannot freely rotate within the binding site, and it preferentially unbinds and rebinds to interconvert between the two bound states.

The previous study-cases sampled only one unbinding event for each replica and, for this reason, the graph model only describes the interconnection between states along the unbinding pathway. In this last case, due to the MetaD sampling of several binding/unbinding events, the obtained graph takes into account connections along both directions, and thus contains more information about the kinetic of the process. Indeed, assuming that (in the limit of a quasi-equilibrium process) the ligand remains trapped for a sufficient time inside an energy minimum, the communities identified with the walktrap method exhibits the properties of kinetic clustering (Figure 4.16b). The identified communities well represent the ensemble of metastable states sampled along the process. Along both branches it is possible to identify: a small community for the bound state (communities E and G); a community in which the ligand is still completely inside the binding cavity and did not reach the unbound state (C and A);

a community in which the ligand is located at the mouth of the cavity, but it is already partially immersed in the solvent (B and D); a community for the completely unbound state (F). Moreover, the transitions between communities may be associated to conformational changes with high energy barriers. Focusing on transitions between communities B and C, and between communities A and D, it seems that they are associated to the conformational changes necessary to observe ligand binding. Indeed, nodes at the boundary of these two pairs of communities display higher average RMSD values for residues at the mouth of the cavity involved in the recognition process (Figure 4.17).

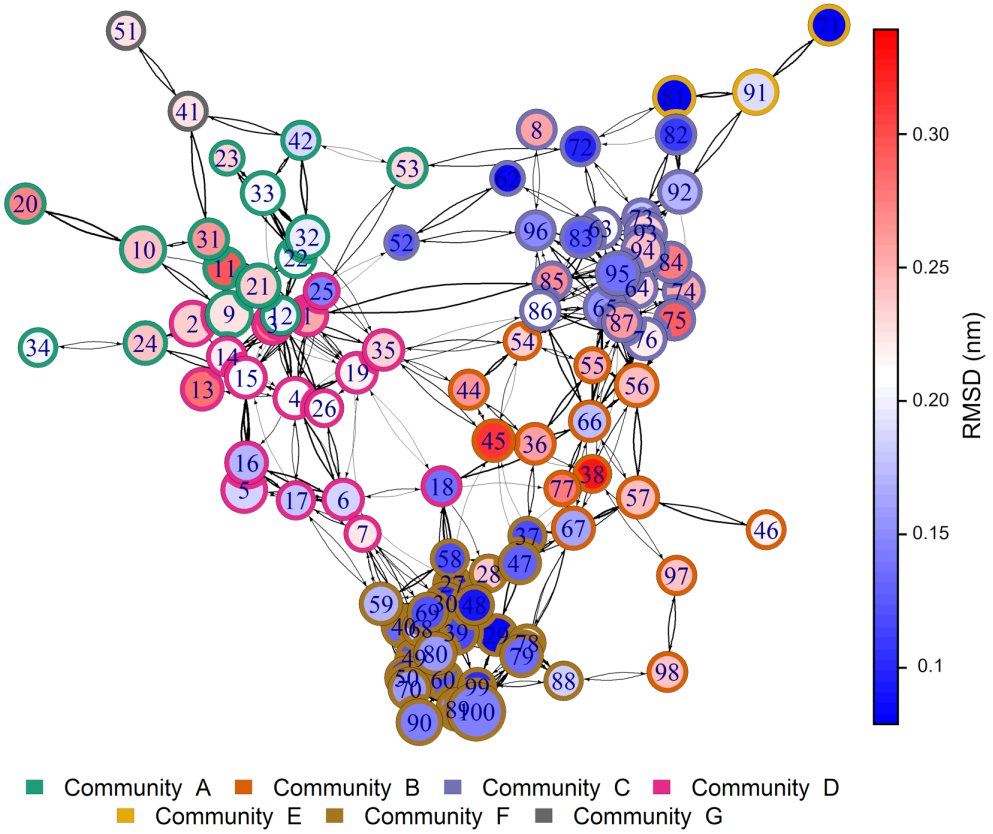

Figure 4.17 Transition network for the MetaD simulation of THS-020 binding to HIF-2a. Nodes are colored according to the average RMSD value of residues belonging to the mouth of the cavity (computed on backbone atoms of residues 288-293 and residues 302-306). The circle around each node is colored according to the community it belongs to.

Finally, we performed a committor analysis: we computed the probability of ending in the crystallographic-like bound conformation (neuron 91, in the community E) before reaching the unbound conformation (neuron 100, community F) starting from each neuron (Figure 4.16c). Given that the transition state is expected to have an equal

chance of going to either states, configurations with a committor of approximately 0.50 can be considered at the transition state. In the present case, the energetic barrier seems to be located around conformations close to neurons 63, 72, 73 and 82 (in the community C, Figure 7c). These conformations are located at the boundaries between communities E and C and are near to the bound state, in agreement with the conclusions drawn from the sMD simulations.

4.5 Conclusions

Data from MD simulations can contain extremely useful information on molecular processes, but it does not lead to simple canonical analysis protocols: system-specific and problem-specific strategies are often required to extract information from increasingly large trajectory files. Planning and designing appropriate strategies can be a very difficult task, and it often requires the development of ad-hoc scripts for advanced analysis and the use of dedicated analysis tools.

Several general-purpose tools for the analysis of MD trajectories are available, including GROMACS analysis tools[175], CPPTRAJ[176], VMD[133], MDAnalysis[177], Bio3D[178] and MDTraj[179]. All these tools provide basic post-processing analysis such as RMSD, RMSF, radius of gyration, hbond and contact maps. Some of them are built-in tools distributed along with the main simulation engine (GROMACS analysis tools and CPPTRAJ), while others are python or R libraries that provide a flexible framework for complex analysis (MDAnalysis, Bio3D and MDTraj), but require the user to develop ad-hoc code.

Among the most advanced post-processing methods, Markov State Models (MSMs) are often used to develop a complete kinetic model of the process under investigation[142,180]. These types of analysis are often complex and require a high level of expertise by the user to obtain reliable results. For this reason, they are difficult to implement as an automated user-friendly protocol. Moreover, effective use of MSMs requires that simulated data meet strict sampling conditions, such as a lag time sufficiently long to produce a Markovian state decomposition[181]. This implies that this

method can only be used when the aggregate simulation time is in the order of hundreds of microseconds or more. Moreover, development of MSMs using enhanced sampling MD requires reweighting procedures that nowadays are still at an early stage of development[182].

Here we presented a tool based on SOMs specifically designed for the analysis of ligand binding pathways sampled in simulations by means of an automated protocol. Our development takes inspiration from other tools based on SOMs already developed by our group and others[31], some of which with fast implementation on GPU[30]. These tools have been successfully applied to MD data, but they were mainly focused on clustering of macromolecular conformations and not on pathway analysis.

The PathDetect-SOM tool does not have any sampling condition and can be applied to MD simulations that sample multiple ligand-binding events. While it cannot be directly used to compute stationary quantities and long-time kinetics (unless one demonstrates that the criteria for MSMs are met), it provides an immediate interpretation of the pathways sampled during the simulation, and can give hints about the thermodynamics and kinetics of the process.

In this work, we tested the tool on a range of ligand binding/unbinding simulations with different features. In all cases the pathways were successfully characterized and mapped over an intuitive 2D map, thus confirming the general applicability of the protocol. Moreover, depending on the simulation type, several hints regarding the energetics of the process were obtained. In the first study-case, we exploited the possibility of re-mapping a property, the sMD pulling forces, on the SOM neurons in order to identify the location of the highest unbinding energy barrier along the simulation (corresponding to the frames with the largest values of the pulling forces). In the second study-case, the transition graph and the betweenness centrality score of the nodes suggested the obligate transition across a neuron for the unbinding across pathway 1. Finally, in the third study-case, we treated the simulation as a quasi-equilibrium simulation and computed some interesting properties starting from the approximate transition matrix. The committor analysis suggested the location of the energy barrier on the SOM, while determination of the communities in the transition

graph led to the identification of kinetic macrostates. As the above properties were computed from the approximate transition matrix, their accuracy strictly depends on the extension of the sampling.

PathDetect-SOM has been implemented in the form of an R batch script with an easy command line interface. While the tool was primarily designed for ligand binding studies, it can be applied to many other types of simulations (unfolding, protein-protein or protein-peptide binding) by appropriate choice arguments on the command line input. The batch script format offers easiness of use with flexibility of customization through simple command line options. As future development the tool can be extended and included in an R package to offer expert users the possibility to develop ad-hoc extensions to the analyses. The tool is open source and freely available with a brief guide and tutorials at https://github.com/MottaStefano/PathDetect-SOM.

INVESTIGATION OF LIGAND-PROTEIN INTERACTION THROUGH HIGHLY SCALABLE QM/MM MD SIMULATIONS

5.1 Introduction

Predicting energetics and kinetics of ligand-protein interactions is crucial for both basic science and applications. From a biophysical perspective, the molecular recognition process of small molecules interacting with protein is one of the fundamental biochemical processes such as metabolic pathways and neurotransmission. In drug design, the affinity[183] and the residence time[184] are key quantity that define the efficiency of a drug binding to its target protein. While the calculation of the first is well established and it benefits to a plethora of powerful approaches (including those discussed in Chapters 1 and 2), the prediction of kinetics quantities (in particular of ligands' k_{off} values) still poses challenges. Important studies from the Parrinello's[64,185,25,62,131 186] and Noe's[187,188,189] groups have shown the feasibility of advanced computational methods such as Infrequent Metadynamics[26] (InMetaD) and Markov State Models[180,181] (MSMs) to calculate k_{off} values (or residence times).

Unfortunately, however, while in some cases such predictions have been relatively accurate[131], in other cases[190] significant discrepancies with experiment have been found. This has been ascribed, at least in part, to errors of the force field in describing the energetics of the transitions state of the binding/unbinding events[190]. This is not surprising, as force fields have been not parametrized so as to reproduce structure and energetics of such transition states.

First principles based QM/MM approaches to ligand unbinding could help address this issue. Here, a part of the system is treated at the quantum level (QM part), for instance using density functional theory (DFT) while the rest of the system is handled by a classical force field (MM part). These approaches allow for a dramatical decrease in the size of the computationally expensive QM part, while retaining the ability to describe specific quantum mechanical processes (such as enzymatic reactions[191,47], proton transfer phenomena[192]) using quantum chemical methods such as density functional theory (DFT). In the context of ligand unbinding, QM/MM might be able to describe with not too dissimilar accuracy all the configurations explored by ligand binding (including the transition state). Machine learning approaches such as those developed in ref[193] could then be used to greatly enhance the convergence of the QM/MM calculations.

Recently, the Carloni's group in Jülich, in collaboration with an European network including EPFL, proposed a flexible and efficient QM/MM that aims to achieve unprecedented scaling in QM/MM simulations[194]. This scheme is called MiMiC: multiscale modeling in computational chemistry and it can cover subns timescales using DFT for the quantum part[195]. Here I explored the use of such massively parallel QM/MM simulations to investigate ligand binding/unbinding events. Because of restrictions of time, we focused on the first step of the latter, namely dynamics of the ligand bound to its binding site.

We performed as many as 40.3 ps of QM/MM simulations the ligand 2g (Figure 5.4a) bound to the mitogen-activated protein kinase p38 (Figure 5.3). The latter is a very important pharmaceutical target for which excellent computational kinetic

studies have been already performed[65,196,197,198,199] . The QM part consisted of the ligand and was treated at the DFT-BLYP level.

Here we derive some key electronic properties of ligand/protein dynamics, such as the electronic polarization of the ligand. These properties are impossible to obtain by standard, force field-based MD simulations. To the best of our knowledge, this is the first time that a ligand-protein system is studied by MiMiC QM/MM MD. Indeed, while simulations on simple systems[200,194,201] and membrane proteins (anion channels)[192,195] have already appeared (showing the impressive scaling of the code[202]), QM/MM simulations of ligand/protein complexes have been limited to single point calculations[203]. Thus, although I focus on only one step of the unbinding process, I do provide here the first dynamics study of a protein in complex with its ligand using this new, powerful QM/MM code.

MiMiC: Multiscale Modeling in Computational Chemistry

MiMiC is based on a multiple-program multiple-data (MPMD) model with loosely coupled programs (Figure 5.1): a main driver running molecular dynamics simulations coupled to a set of external programs, each of which computes contributions that are relevant to a specific subsystem using their optimal parallelization strategies. This strategy allows the use of different models such as QM, MM, CG (coarse-grained), CM (continuum mechanics). Communication between programs is possible by using a lightweight communication library (CommLib)[200].

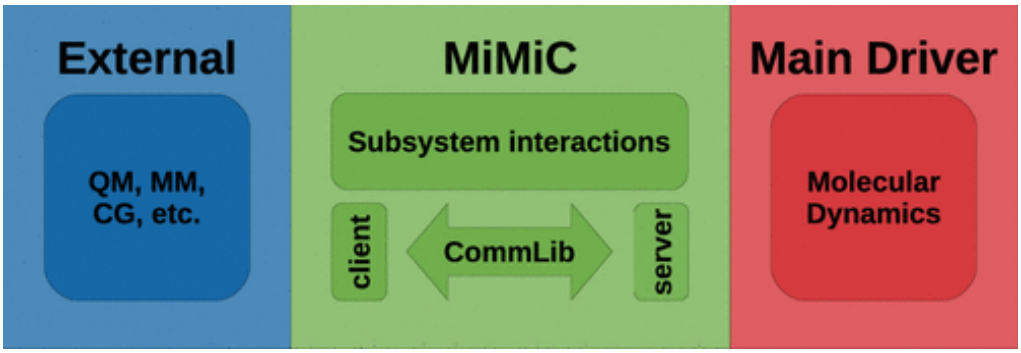

Figure 5.1 Schematic representation of the strategy used in the MiMiC framework (Image from ref [200])

In the present implementation, MiMiC[201] couples two highly efficient programs CPMD[204] that serves ad main driver and also computes the QM contribution, and

GROMACS[117] that provides the MM contribution. The CPMD program uses a plane wave/pseudopotential implementation of DFT, and it allows to perform both Born-Oppenheimer (BO) and Car-Parrinello (CP) QM/MM MD simulations. The workflow of a simulation employing the MiMiC framework is illustrated in Figure 5.2. Both programs are run independently and simultaneously using MiMiC both to communicate and to calculate QM/MM contributions. In the initialization phase, GROMACS and CPMD read their respective input files, and MiMiC collects data from both and sends the necessary data to CPMD, e.g. coordinates, atom types, etc. At this point, the QM/MM-MD cycle is entered; it consists of several steps: the first is to send the coordinates to GROMACS which then proceeds to calculate the energy and MM forces. At the same time, CPMD calculates the corresponding QM contributions subject to the electrostatic potential calculated by MiMiC on the QM grid. MiMiC also calculates the QM/MM energy and forces. Finally, all force contributions are collected, and CPMD integrates the equations of motion and continues to the next iteration of the MD loop[200].

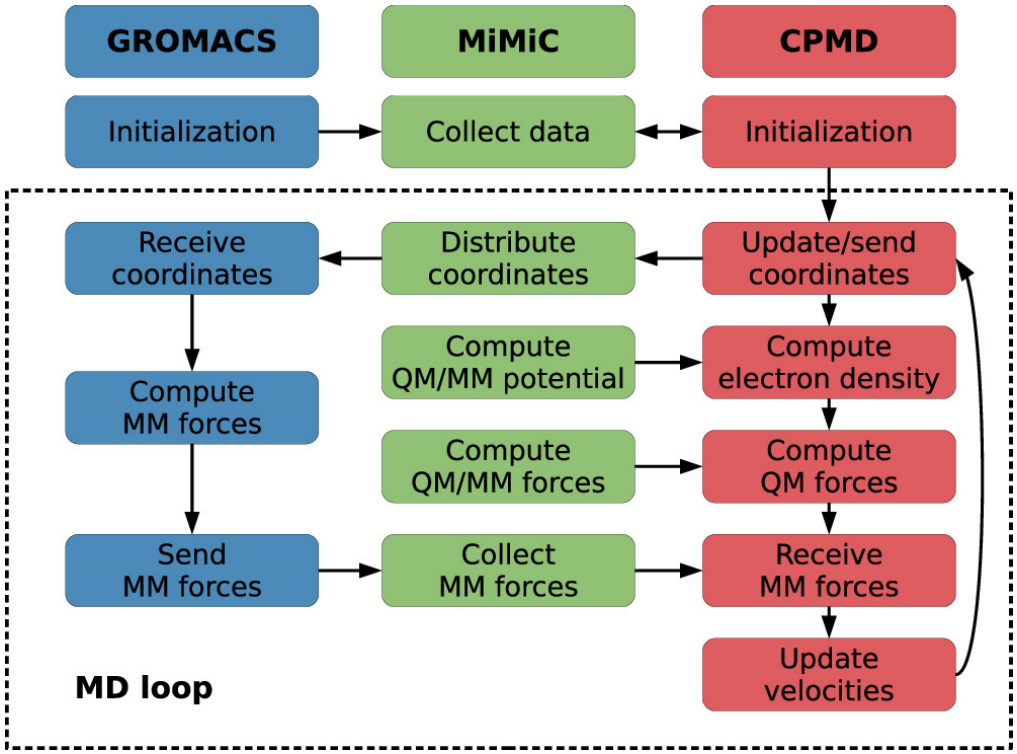

Figure 5.2 Schematic representation of QM/MM MD workflow using the MiMiC framework (Image from ref [200]).

Study-case: Mitogen-activated protein kinase p38 in complex with 2g ligand

The mitogen-activated protein kinase p38 is a member of the mitogen-activated protein kinase (MAPK) family. It is a serine/threonine kinase that controls cytokine biosynthesis and is involved in the initiation of chronic inflammation processes, development of cancer, heart disease, and many others[205,206]. Four different isoforms (α, β, γ, δ) of p38 MAPK family have been identified. Several genes encode them: p38α (MAPK14), p38β (MAPK11), p38γ (MAPK12, and p38δ (MAPK13). P38 MAPK adopts a typical kinase fold, including N-terminal lobe, rich in β -sheet (blue in Figure 5.3), and C-terminal lobe (grey in Figure 5.3), rich in α-helix, that are connected via a hinge region (light green in Figure 5.3). The catalytic site of the protein is placed between the two lobes, where ATP molecules bind. The ATP binding site of p38 is composed by the hinge region, the glycine-rich loop (orange in Figure 5.3), the activation loop (magenta in Figure 5.3), the Asp168-Phe169-Gly170 amino acids which compose the DFG motif (yellow and cyano in Figure 5.3), and the αC-helix (dark green in Figure 5.3). Between the two lobes there is also an allosteric site[207] (AS) that is created by the movement of the DFG motif between two conformations: the active conformation (DFG-in, cyano in Figure 5.3) and the inactive conformation (DFG-out, yellow in Figure 5.3). When the DFG motif is in its “in” conformation, the AS is filled by the Phe169 side chain, while, when the DFG is in the “out” conformation the Phe169 is placed on the opposite site preventing the ATP binding[206].

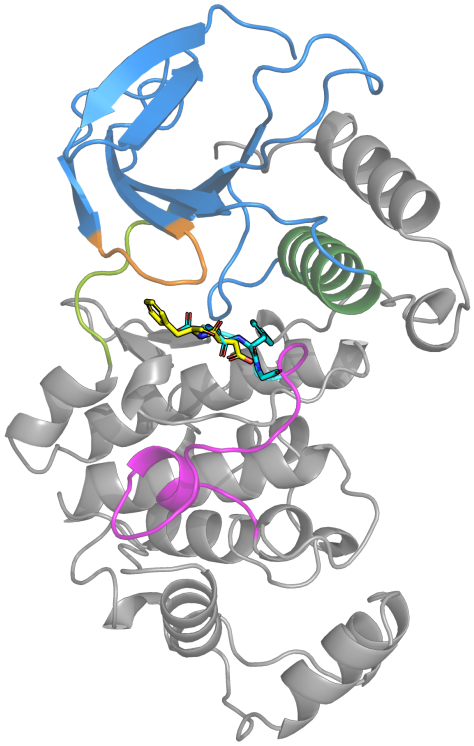

Figure 5.3 3D structure of p38 MAPK: protein is shown as cartoon, the N-terminal lobe in blue, the C-terminal in grey, the glycine-rich loop in orange, the activation loop in magenta, the αC-helix in dark green. The DFG motif is shown in sticks: DFG in its "in" conformation in cyano and DFG in its "out" conformation in yellow.

All the isoforms contain a conserved dual phosphorylation motif and both phosphorylations are necessary to fully activate the kinase. Dual phosphorylation at these sites alters the folding of p38 by stabilizing the activation loop in a more open conformation and causing rotation between the two lobes, which allows substrate recognition and increases the activity of the kinase. The p38α isoform is one of the most studied kinases as it has been identified as a drug target for various diseases and several inhibitors have been proposed over the years[206]. Among all available inhibitors for the α isoform of p38, the 2g ligand ((6-(2-fluorophenoxy)-8-methyl-2-(tetrahydro-2H-pyran-4-ylamino)pyrido[2,3-d]pyrimidin-7(8H)-one), in **Error! Reference source not found.**a) was chosen for this work. The structure of a very close analogue, ligand 2a, (6-(2,4-difluorophenoxy)-8-methyl-2-(tetrahydro-2H-pyran-4-ylamino)pyrido[2,3-d]pyrimidin-7(8H)-one, in Figure 5.4b) in complex with the p38 protein has been determined by X-ray crystallography.

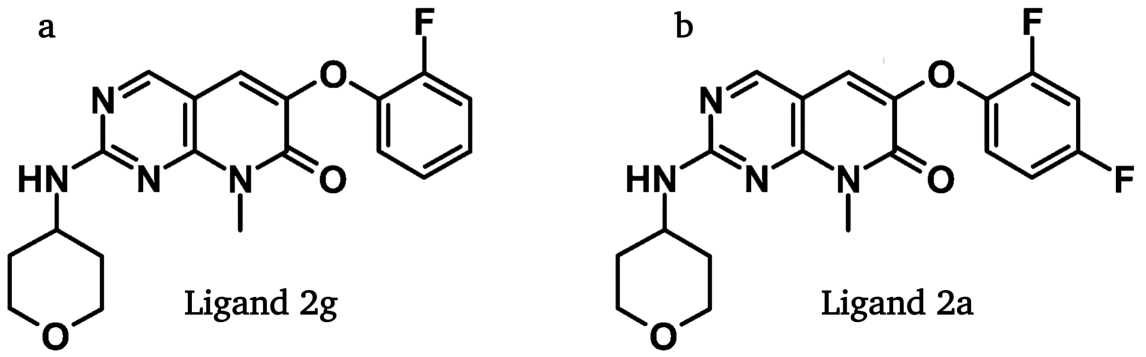

Figure 5.4 2D structure of the 2g (a) and 2a (b) ligands The ligands are the same except that a hydrogen is replaced by a fluorine.

The p38-2g complex is part of a dataset[208] published in conjunction with the results of a high-throughput FEP workflow developed by Gapsys and coworkers[209].

5.2 Methods

Force field based MD simulations on the complex in aqueous solution were carried out by my colleague Katya Ahmad in Jūlich and they are summarized in Appendix at the end of this chapter.

The QM/MM MD simulations were performed using the snapshot obtained after the 150 ns of NPT equilibration from the force field-based simulations. We used the MiMiC framework, coupling the CPMD[204] 4.3 version and the GROMACS[117] 2019.4 version. The code ran on the clusters hosted by Jülich Supercomputing Center namely JURECA[210] and JUWELS[211]. Within MiMiC, a generalized version of the electrostatic embedding scheme introduced by Laio et al[212] was used, in which the total energy of the system is calculated following an additive scheme. The QM/MM simulations were performed with a DFT approaches based on the use of plane wave as basis set and pseudopotentials. Specifically, all the simulations were performed using Troullier-Martins norm-conserving pseudopotentials[212]. The system consists of 169,550 atoms, were 46 atoms (i.e. the atoms of the ligand) were selected to represent the QM subsystem. This was treated at the BLYP[213,214] level, while the MM is described by the Amber99SB*-ILDN[215,216] force field. 40.3 ps of Born-Oppenheimer MD/force field based MD NVT simulations were carried out using a time step of 0.48 fs. Constant

temperature simulations were achieved using the Nose-Hoover thermostat[217,218] with a coupling frequency of 5000 cm^{-1}.

Analysis

The final frames of the ten production mode MM simulations and nine snapshots (one every 5 ps plus the initial one) from the QM/MM simulation were extracted for analysis.

Structural features: A visual analysis of H-bonds and hydrophobic interactions were performed with Protein-Ligand Interaction Profiler[219], comparing the frames derived from the simulations and the X-ray structure of the analogous ligand, 2a, which replaces a hydrogen atom with a fluorine (see Figure 5.4).

Electronic properties: The electronic density of the ligand (both in vacuo and in the bound state) was calculated at the BLYP level using 6-31G(d,p)[220,221] as basis set for each QM/MM snapshot, using the Gaussian[222] program. The electric field of the protein and of the solvent was introduced in some of the calculations.

The change in electron density was computed as:

$$\Delta\rho = \rho_{lig}^{complex} - \rho_{lig}^{vacuo}$$

The density of the complex is obtained by performing the calculation in the presence of the electric field of the surrounding protein and the aqueous solvent. By integrating the $\Delta\rho$ it is possible to monitor the change in atomic charge for each ligand atom (i):

$$\Delta Q(i) = \int \Delta\rho(r)dr$$

The integral is solved numerically over the grid points within the Voronoi[223] partition of atomi i (VP_i) using the code from ref[190]. An estimation of the change in charge distribution is given by electric polarization as:

$$\Delta Q_{Pol} = |\Delta Q(+)| + |\Delta Q(-)|$$

where:

$\Delta Q(+) = \sum_i \Delta Q(i), i \in \{\Delta Q(i) > 0\}$ and $\Delta Q(-) = \sum_i \Delta Q(i), i \in \{\Delta Q(i) < 0\}$

5.3 Results

In this work, the classical MD simulations are solely used to equilibrate the system before performing the QM/MM calculations. Selected results are reported in Appendix at the end of this chapter.

Structural analysis. Our QM/MM simulations are consistent with the experimental structural information and provide insight on the interaction of the ligand with the solvent in the complex in aqueous solution.

The H-bonds formed by the protein with the ligand are reproduced (Table 5.1). Novel insight is obtained on the hydration of the ligand: the atoms O1 and O2 form H-bonds with water, while the H-bond of the water molecule bridging the water and the protein is still present. However, this water interacts here also with another water molecule from the solvent, at variance of the crystal structure. The hydrophobic interactions are also maintained, and, in some cases, strengthened (Figure 5.5 and Table 5.2)

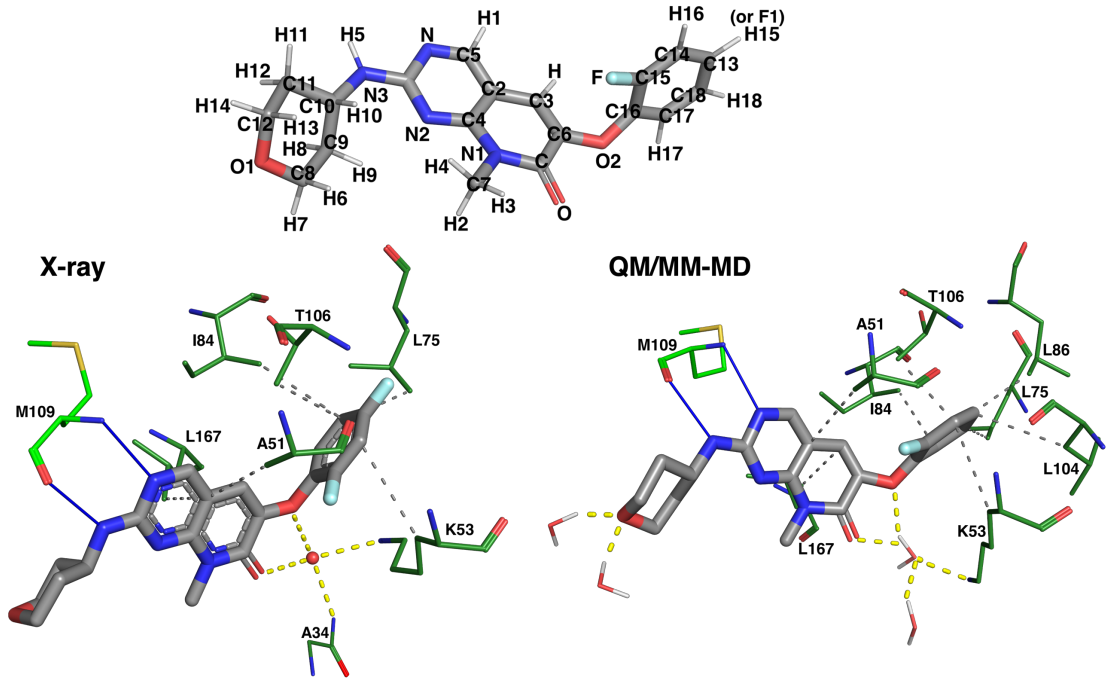

Figure 5.5 3D representation of selected ligand-protein interactions. Top: 3D representation of the ligand with atom labelling. Bottom, left: ligand-protein interactions observed in the X-ray structure; bottom, right: ligand-protein interactions observed across nine QM/MM snapshots. The ligand is shown in sticks and the residues involved in the interactions in lines. H-bonds are described by solid blue lines, interaction with water molecules by dashed yellow lines, and hydrophobic interactions by dashed grey lines.

Table 5.1 H-bonds and water interactions observed in the X-ray structure and across the nine snapshots derived from QM/MM simulation (see Figure 5.5). Errors reported as the standard deviation of donor-acceptor distances.

		X-ray distances (Å)		**QM/MM-MD mean distances (Å)**	
Donor	Acceptor	D-A		D-A	D-H-A
N@M109	N@ligand	3.0		2.8 ± 0.1	1.8 ± 0.1
N3@ligand	O2@M109	2.9		3 ± 0.1	2.0 ± 0.1
N@A34/H2O	O2@ligand	2.8	2.8	3.0 ± 0.2	3.6 ± 0.5
NZ@K53	H2O	2.8	3.1	2.9 ± 0.1	2.8 ± 0.1
H2O	O1@ligand	-		2.9 ± 0.6	
H2O	O1@ligand	-		2.5 ± 0.4	

Table 5.2 Hydrophobic interactions observed in the X-ray structure and the nine snapshots derived from QM/MM simulation (see Figure 5.1). Errors reported as the standard deviation of donor-acceptor distances.

		X-ray	**QM/MM-MD**
Ligand atom	Protein atom	Distance (Å)	Mean Distance (Å)
C2	CB@A51	3.7	3.7 ± 0.2
C14	CB@K53	3.9	3.6 ± 0.2
C18	CD2@L75	3.8	3.8 ± 0.2
C17	CG2@I84	3.9	3.7 ± 0.2
C13	CD2@L86	7.1	3.7 ± 0.2
C13	CB@L104	4.8	3.8 ± 0.1
C14	CG2@T106	3.8	3.4 ± 0.2
C2	CD2@L167	4.0	3.7 ± 0.1

Electronic Properties. We computed, using QM/MM, the rearrangement of the electronic density of the ligand as it passes from *vacuo* to the bound state. The calculations are carried out using the Voronoi partition[199] of the atomic charges.

The Figure 5.6 shows the change in electronic density of the ligand on passing from in vacuo to the bound state for selected QM/MM snapshots. The Figure 5.7 plots the corresponding change in electronic charge, $\Delta Q(i)$ for each atom of the ligand during

the 40.3 ps QM/MM dynamics. Here, ΔQ(+) and ΔQ(-) which are the polarizations of the ligand contributed by atoms with positive and negative ΔQ(*i*), respectively. They range from -0.064 to 0.056 electrons. Thus, polarization effects are small but not negligible. A similar conclusions was reached by Capelli et al[190] for a ligand-receptor complex, although in that case only single points were calculated. Importantly, the values were not constant over time (Figure 5.7); this confirms the importance of conformational fluctuations for accurately describing ligand/protein electrostatic interactions. The values of the charges and their fluctuations are expected to change during the unbinding process. Including these variations of the charges might be very important to describe protein/ligand interactions, cannot be captured by standard force field-based MD simulations.

The largest variation observed (in absolute value) is that of the pyrimidine nitrogen (N, Figure 5.5) which forms an H-bond as acceptor (Table 5.1). Several other atoms forming H-bonds either with the protein or the solvent experience sizeable polarization effects. Not unsurprisingly, the latter are far less pronounced in the hydrophobic portion of the ligand (Figure 5.5, bottom right), which establishes hydrophobic interactions within the binding cavity with residues A51, K53, L75, I84, L86, L104, T106, and L167 (Table 5.2).

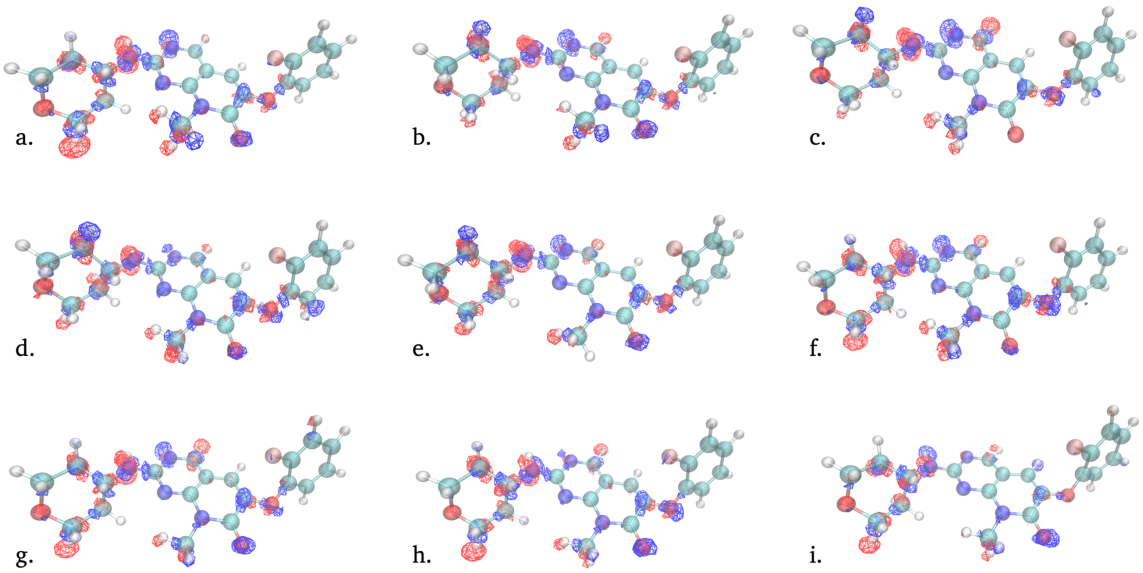

Figure 5.6 Change of electronic density of the ligand on passing from vacuo to bound state (Δρ) for 9 different snapshots (from 0 (a.) to 40 (i) ps every 5 ps) during the QM/MM simulations. The ligand atoms are displayed as stick, the electronic difference as isosurface, blue Δρ >0 and red Δρ <0.

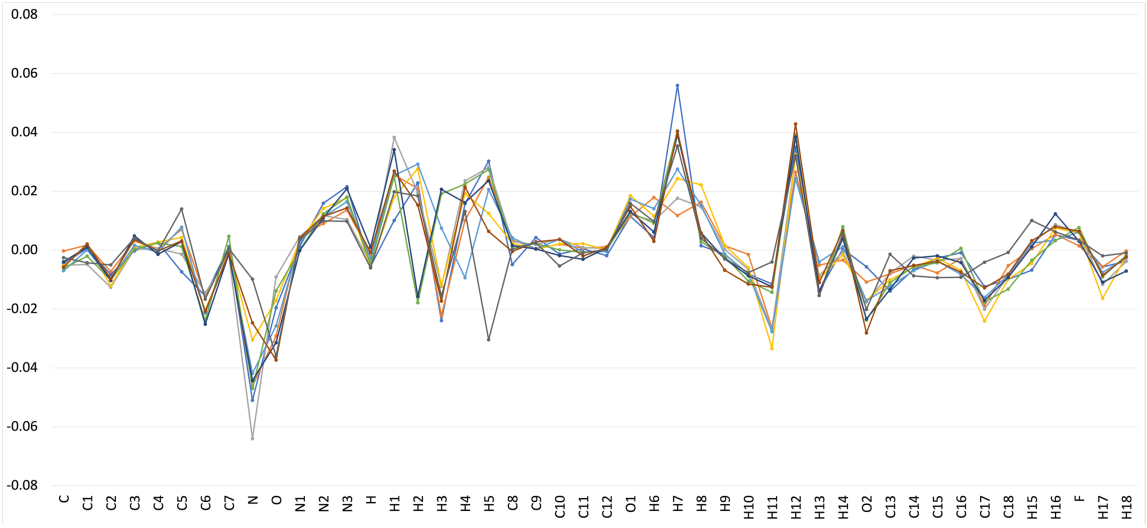

Figure 5.7 Change in atomic charge for each ligand atom (ΔQ(i), i=0-45) across the nine QM/MM snapshots.

5.4 Conclusions

Our QM/MM calculations reproduce the pose of the X-ray structure and provide insights on polarization effects of the protein electric field onto the ligand. These effects are small albeit significant. They are mostly pronounced on the pyrimidine nitrogen, which forms H-bond as an acceptor. Interestingly, they vary significantly during the dynamics, pointing of the relevance of conformational fluctuations for ligand/protein electrostatic interactions. We can expect that overall, these effects may vary during the unbinding process as the ligand interact with different groups in the process. The energetics associated with these small albeit significant changes of interactions cannot be captured by standard biomolecular force fields. Also charge transfer effects, not investigated here, could play a role for ligand/protein interactions[190].

Appendix: Force field based MD simulations

Methods

The initial structure and the topology file of p38-2g complex were taken from ref[208]. The topology contained Amber99SB*-ILDN force field parameters for the protein and GAFF parameters for the ligand. The initial structure of the protein-ligand (2g) complex from the dataset[208] was stripped of the pre-existing water and ions and solvated in a 12x12x12 nm cubic box of TIP3P water and neutralized with Na^+ and Cl^- ions corresponding to a concentration of 0.033 mol dm^{-3} of NaCl. The final system was subjected to 3000 steps of energy minimization through the steepest descent algorithm implemented in GROMACS[117] 2020.4. All further equilibration and production simulations were conducted using a 2 fs time step, P-LINCS[224] algorithm for imposing constraints, the Bussi velocity-rescaling thermostat[225] with a time constant of 0.1 ps, and PME[123] electrostatics with a 1 nm cutoff and a Fourier grid spacing of 0.12 nm. After energy minimization, the system was equilibrated for 1 ns in the NVT ensemble with the temperature maintained at 300K. This was followed by 1 ns of NPT equilibration at 1 bar with the Berendsen barostat[120] with a time constant of 2 ps for the initial relaxation of the box volume, followed by a further 150 ns of equilibration at 1 bar with the Parrinello-Rahman barostat[122], with a time constant of 2 ps. The duration of this equilibration phase was sufficient for the RMSD of the protein to converge at 0.208 ± 0.027 nm. An ensemble of ten production mode force field MD simulations of 50 ns duration were then conducted.

Selected results

In this work, the classical MD simulations are solely used to equilibrate the system before performing the QM/MM calculations. Selected results from MD, provided by my colleagues Katya Ahamad, suggest that the MD reproduces the ligand binding pose. Indeed, (i) The main H-bonds interactions of the ligand with the protein are fully maintained during the dynamics (Table 5.3); (ii) the hydrophobic interactions are also rather well maintained.

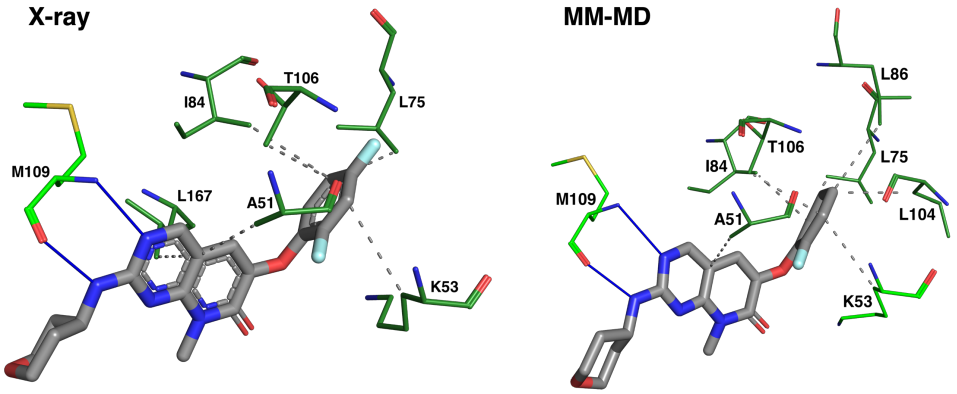

Figure 5.8 3D representation of selected ligand-protein interactions. Bottom, left: ligand-protein interactions observed in the X-ray structure; bottom, right: ligand-protein interactions observed across the final frame of the MD simulation. The ligand is shown in sticks and the residues involved in the interactions in lines. H-bonds are described by solid blue lines and hydrophobic interactions by dashed grey lines. Labelling of ligand atoms as in Figure 5.5.

Table 5.3 H-bonds observed in the X-ray structure and across the final frames of ten replicas of MM production (see Figure 5.8). Errors reported as the standard deviation of donor-acceptor distances.

		X-ray	**MM-MD**	
Donor	Acceptor	D-A distance (Å)	Mean D-A distance (Å)	Mean D-H-A distance (Å)
N@M109	N@ligand	3.1	3.1 ± 0.1	2.1 ± 0.2
N3@ligand	O2@M109	2.9	3.0 ± 0.2	3.0 ± 0.3

Table 5.4 Hydrophobic interactions observed in the X-ray structure and across the final frames of ten replicas of MM production (Figure 5.8). Errors reported as the standard deviation of donor-acceptor distances.

		X-ray	**MM-MD**
Ligand atom	Protein atom	Distance (Å)	Mean Distance (Å)
C2	CB@A51	3.8	3.7 ± 0.1
C14	CB@K53	3.9	3.7 ± 0.2
C18	CD2@L75	3.8	3.7 ± 0.2
C17	CG2@I84	3.9	3.7 ± 0.2
C13	CD2@L86	7.1	3.8 ± 0.1
C13	CB@L104	3.8	3.6 ± 0.2
C14	CG2@T106	3.8	3.5 ± 0.1
C2	CD2@L167	3.9	7.8 ± 0.1

CONCLUSIONS

In this PhD thesis I focused on modeling of ligand-protein binding with computational methods based on molecular dynamics. Understanding this process is crucial for the design and discovery of new drugs and the use of computational methods to support experimental research in this field is constantly growing.

Over the years, many computational methods[7] have been developed to address the study of ligand-protein binding. Nowadays, methods based on classical molecular dynamics have become central to this field thanks to the increase in computing power. In particular, in this thesis I made use of physical pathways (PP) methods based on enhanced sampling techniques[14]. This type of methods enables the simulation of the complete binding and/or unbinding events, allowing not only to estimate the ligand-protein binding affinity, but also to obtain information about the ligand binding pathways[14]. However, the use of these methods requires the production of several replicas or long simulations to sample the binding/unbinding event several times in order to obtain a reliable statistics of the process. This produces the need of methods able to analyze all the simulated events at once and to provide a clearly interpretable picture of the differences in the sampled pathways. On the other hand, the use of

methods based on MM still suffer from several limitations due to the limited accuracy of force-fields in the description of ligand-protein interactions. QM approaches can indeed provide a better description of the system; however, as well known, performing QM simulations on large biomolecular systems requires prohibitive computational costs. Hybrid QM/MM approaches are useful in this perspective, given that they allow to obtain a more accurate description of the process while retaining low computational costs[47].

In this thesis, an approach based on the combination of some efficient enhanced sampling methods, the steered MD (sMD) and the Metadynamics (MetaD), was proposed[108]. It was applied to predict the possible binding/unbinding pathways of some ligands to the HIF-2α PAS-B domain and to obtain a correct estimation of their binding free energy (chapter 3). Modeling these processes represents a non-trivial task due to the buried nature of the binding cavity that suggests significant protein conformational changes may occur upon ligand access. Indeed, the proposed computational protocol was successful in modeling these events. In fact, with sMD it was possible to select the preferred unbinding pathway, by performing multiple replicas in parallel and thus comparing different pathways with relatively low computational costs. In addition, the use of MetaD simulations allowed to overcome some sMD limitations regarding the correct estimation of the binding free-energy. Some critical points in using MetaD simulations needed to be addressed: the first one concerns the choice of collective variables suitable to correctly describe the process, the second one is the construction of a correct reference path. The first point was addressed by employing the Path Collective Variables approach, while, for the second point, simulations obtained with sMD were adopted to derive the reference path. In this way it was possible to estimate the correct binding free-energy values for the analyzed ligands as well as to characterize the relevant states along their unbinding pathways. The protocol here proposed appears to be an invaluable tool to investigate the binding process of different ligands, using simulations performed both on a known ligand-protein X-ray structure and on a docking pose, thus contributing to the development of successful drug design campaigns. In addition, the results obtained

encourages to extend its application to other binding mechanisms involving systems with characteristics similar to those of our study-case.

As mentioned above, to address the difficulty of analyzing large amounts of data derived from several replicas (as in the case of sMD) or from a single long simulation (like in MetaD), a tool based on the self-organizing maps (SOMs) was proposed[173] (chapter 4). The PathDetect-SOM (Pathway Detection on SOM) tool, has proved to be effective in the analysis of ligand binding/unbinding pathways derived from MD simulations with PP methods. Its general applicability was demonstrated by addressing some study-cases whose simulations had been performed with different methods and thus exhibit different characteristics. In particular, the first study-case (sMD) explored the simultaneous evolution of the replicas (due to the constant velocity of the bias) and the use of a directional Collective Variable (CV). Differently, the method used for the second study-case (Infrequent Metadynamics) involves a type of non-directional CV and may provide very different unbinding paths. Finally, in the third study-case the MetaD simulation evolves in all the directions along two selected CVs and the ligand has greater freedom than in the previous cases. The tool made it possible to analyze multiple simulated events at the same time and to provide a clearly interpretable overview of the differences in the sampled pathways. In addition, hints on the kinetic and thermodynamic properties of the analyzed processes were derived. While the tool was here designed for ligand binding studies, it can be applied to many other types of simulations (unfolding, protein-protein or protein-peptide binding) by appropriate choice of input arguments.

Finally, as part of a project in collaboration with the Prof. Paolo Carloni, the first step of the study of the unbinding process of the mitogen-activated protein kinase p38 in complex with the 2g ligand was explored through the use of QM/MM simulations with the MiMiC interface. Our DFT-based QM/MM simulations, carried out for a relatively long trajectory (more than 40 ps) exploiting a highly scalable interface, allowed us to describe the dynamics and the electronic properties of the p38-2g complex. Our calculations suggest that polarization effects are small, although significant. They are expected to change during the dynamics, as observed in the case

of another receptor/ligand complex (although in that case only static calculations were carried out)[190]. In the latter system, these effects turned out to be relevant for the energetics of protein/ligand interactions. They are therefore expected to impact on the energetics of the unbinding process (and thus on the kinetics) also in the enzymatic system investigated in this thesis. Of course, running QM/MM simulations for the entire unbinding process is unfeasible with current resources, even exploiting an extremely scalable code as that used here and large supercomputers such as those in Jūlich. However, recent advances in machine learning-based predictions of free energy calculations based on QM and QM/MM methods[193] suggest that in a not too distant future, highly scalable QM/MM simulations, along with machine learning approaches, contribute to accurate descriptions of unbinding processes towards the quantitative prediction of ligand residence times across a wide variety of systems.

In conclusion, in this thesis I faced the problem of the computational study of ligand-binding from several perspectives. The definition of an enhanced sampling protocols for the study of the binding pathways, the development of a tools for the analysis of PP simulations, and an improved description of protein-ligand interactions though QM/MM approaches, are all open challenges that aim to increase the accuracy of calculations. With the increasing computational power, such methods will become routinely used and could be determinant in the success of a drug design campaigns.

REFERENCES

1. Babine, R. E. & Bender, S. L. Molecular Recognition of Protein−Ligand Complexes: Applications to Drug Design. *Chem. Rev.* **97,** 1359–1472 (1997).
2. Baron, R. & McCammon, J. A. Molecular recognition and ligand association. *Annu. Rev. Phys. Chem.* **64,** 151–175 (2013).
3. Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science (80-.).* **303,** 1813–1818 (2004).
4. Leelananda, S. P. & Lindert, S. Computational methods in drug discovery. *Beilstein J. Org. Chem.* **12,** 2694–2718 (2016).
5. Du, X., Li, Y., Xia, Y. L., Ai, S. M., Liang, J., Sang, P., Ji, X. L. & Liu, S. Q. Insights into protein–ligand interactions: Mechanisms, models, and methods. *Int. J. Mol. Sci.* **17,** 1–34 (2016).
6. Copeland, R. A., Pompliano, D. L. & Meek, T. D. Drug-target residence time and its implications for lead optimization. *Nat. Rev. Drug Discov.* **5,** 730–9 (2006).
7. 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).
8. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161,** 269–288 (1982).
9. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discov.* **3,** 935–949 (2004).
10. Torres, P. H. M., Sodero, A. C. R., Jofily, P. & Silva-Jr, F. P. Key Topics in Molecular Docking for Drug Design. *Int. J. Mol. Sci.* **20,** 4574 (2019).
11. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and

independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **35,** 539–546 (2010).

12. Lexa, K. W. & Carlson, H. A. Protein flexibility in docking and surface mapping. *Q. Rev. Biophys.* **45,** 301–343 (2012).

13. Buonfiglio, R., Recanatini, M. & Masetti, M. Protein Flexibility in Drug Discovery: From Theory to Computation. *ChemMedChem* **10,** 1141–1148 (2015).

14. Limongelli, V. Ligand binding free energy and kinetics calculation in 2020. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **10,** 1–32 (2020).

15. Åqvist, J., Medina, C. & Samuelsson, J.-E. A new method for predicting binding affinity in computer-aided drug design. "*Protein Eng. Des. Sel.* **7,** 385–391 (1994).

16. Homeyer, N. & Gohlke, H. Free energy calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area method. *Mol. Inform.* **31,** 114–122 (2012).

17. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **10,** 449–461 (2015).

18. Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **3,** 300–313 (1935).

19. Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **22,** 1420–1426 (1954).

20. Bernardi, R. C., Melo, M. C. R. & Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta - Gen. Subj.* **1850,** 872–877 (2015).

21. Izrailev, S., Stepaniants, S., Isralewitz, B., Kosztin, D., Lu, H., Molnar, F., Wriggers, W. & Schulten, K. in *Steered Mol. Dyn.* (eds. Deuflhard, P., Hermans, J., Leimkuhler, B., Mark, A. E., Reich, S. & Skeel, R. D.) 39–65 (Springer, 1999). doi:10.1007/978-3-642-58360-5_2

22. Do, P.-C., Lee, E. H. & Le, L. Steered Molecular Dynamics Simulation in Rational Drug Design. *J. Chem. Inf. Model.* **58,** 1473–1482 (2018).

23. Kästner, J. Umbrella sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1,** 932–942 (2011).

24. You, W., Tang, Z. & Chang, C. A. Potential Mean Force from Umbrella Sampling Simulations: What Can We Learn and What Is Missed? *J. Chem. Theory Comput.* **15,** 2433–2443 (2019).

25. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci.* **99,** 12562–12566 (2002).

26. Tiwary, P. & Parrinello, M. From Metadynamics to Dynamics. *Phys. Rev. Lett.* **111,** 1–5 (2013).

27. Raniolo, S. & Limongelli, V. Ligand binding free-energy calculations with funnel metadynamics. *Nat. Protoc.* **15,** 2837–2866 (2020).

28. Branduardi, D., Gervasio, F. L. & Parrinello, M. From A to B in free energy space. *J. Chem. Phys.* **126,** 054103 (2007).

29. Miao, Y., Bhattarai, A. & Wang, J. Ligand Gaussian Accelerated Molecular Dynamics (LiGaMD): Characterization of Ligand Binding Thermodynamics and Kinetics. *J. Chem. Theory Comput.* **16,** 5526–5547 (2020).

30. Mollica, L., Decherchi, S., Zia, S. R., Gaspari, R., Cavalli, A. & Rocchia, W. Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations. *Sci. Rep.* **5,** 11539 (2015).

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

32. Kokh, D. B., Amaral, M., Bomke, J., Grädler, U., Musil, D., Buchstaller, H. P., Dreyer, M. K., Frech, M., Lowinski, M., Vallee, F., Bianciotto, M., Rak, A. & Wade, R. C. Estimation of Drug-Target Residence Times by τ-Random Acceleration Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **14,** 3859–3869 (2018).

33. Spitaleri, A., Decherchi, S., Cavalli, A. & Rocchia, W. Fast Dynamic Docking Guided by Adaptive Electrostatic Bias: The MD-Binding Approach. *J. Chem. Theory Comput.* **14,** 1727–1736 (2018).

34. Motta, S., Callea, L., Giani Tagliabue, S. & Bonati, L. Exploring the PXR ligand binding mechanism with advanced Molecular Dynamics methods. *Sci. Rep.* **8,** 16207 (2018).

35. Souza, P. C. T., Thallmair, S., Conflitti, P., Ramírez-Palacios, C., Alessandri, R., Raniolo, S., Limongelli, V. & Marrink, S. J. Protein–ligand binding with the coarse-grained Martini model. *Nat. Commun.* **11,** 3714 (2020).

36. Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J. Y., Wang, L., Lupyan, D., Dahlgren, M. K., Knight, J. L., Kaus, J. W., Cerutti, D. S., Krilov, G., Jorgensen, W. L., Abel, R. & Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **12,** 281–296 (2016).

37. Zhang, C., Bell, D., Harger, M. & Ren, P. Polarizable Multipole-Based Force Field for Aromatic Molecules and Nucleobases. *J. Chem. Theory Comput.* **13,** 666–678 (2017).

38. Zhang, C., Lu, C., Jing, Z., Wu, C., Piquemal, J. P., Ponder, J. W. & Ren, P. AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids. *J. Chem. Theory Comput.* **14,** 2084–2108 (2018).

39. Merz, K. M. Using quantum mechanical approaches to study biological systems. *Acc. Chem. Res.* **47,** 2804–2811 (2014).

40. Paquet, E. & Viktor, H. L. Computational Methods for Ab Initio Molecular Dynamics. *Adv. Chem.* **2018,** 1–14 (2018).

41. Hammes-Schiffer, S. & Andersen, H. C. Ab initio and semiempirical methods for molecular dynamics simulations based on general Hartree-Fock theory. *J. Chem. Phys.* **99,** 523–532 (1993).

42. Cheeseman, J. R., Frisch, M. J., Devlin, F. J. & Stephens, P. J. Hartree−Fock and Density Functional Theory ab Initio Calculation of Optical Rotation Using GIAOs: Basis Set Dependence. *J. Phys. Chem. A* **104,** 1039–1046 (2000).

43. Carloni, P., Rothlisberger, U. & Parrinello, M. The role and perspective of ab initio molecular dynamics in the study of biological systems. *Acc. Chem. Res.* **35,** 455–464 (2002).

44. Monard, G. & Merz, K. M. Combined quantum mechanical/molecular mechanical

methodologies applied to biomolecular systems. *Acc. Chem. Res.* **32,** 904–911 (1999).

45. Hayik, S. A., Dunbrack, R. & Merz, K. M. Mixed quantum mechanics/molecular mechanics scoring function to predict protein-ligand binding affinity. *J. Chem. Theory Comput.* **6,** 3079–3091 (2010).

46. Senn, H. M. & Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chemie Int. Ed.* **48,** 1198–1229 (2009).

47. Brunk, E. & Rothlisberger, U. Mixed Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States. *Chem. Rev.* **115,** 6217–6263 (2015).

48. Leach, A. R. in *Mol. Model. Princ. Appl.* 165–252 (2001).

49. Cheng, X. & Ivanov, I. Molecular dynamics. *Methods Mol. Biol.* **929,** 243–285 (2012).

50. Leach, A. R. in *Mol. Model. Princ. Appl.* 353–409 (2001).

51. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23,** 327–341 (1977).

52. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18,** 1463–1472 (1997).

53. 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).

54. Yang, Y. I., Shao, Q., Zhang, J., Yang, L. & Gao, Y. Q. Enhanced sampling in molecular dynamics. *J. Chem. Phys.* **151,** 070902 (2019).

55. Sinko, W., Lindert, S. & Mccammon, J. A. Accounting for Receptor Flexibility and Enhanced Sampling Methods in Computer-Aided Drug Design. *Chem. Biol. Drug Des.* **81,** 41–49 (2013).

56. Lazim, R., Suh, D. & Choi, S. Advances in molecular dynamics simulations and enhanced sampling methods for the study of protein systems. *Int. J. Mol. Sci.* **21,** 1–20 (2020).

57. Patel, J. S., Branduardi, D., Masetti, M., Rocchia, W. & Cavalli, A. Insights into Ligand–Protein Binding from Local Mechanical Response. *J. Chem. Theory Comput.* **7,** 3368–3378 (2011).

58. Patel, J. S., Berteotti, A., Ronsisvalle, S., Rocchia, W. & Cavalli, A. Steered molecular dynamics simulations for studying protein-ligand interaction in cyclin-dependent kinase 5. *J. Chem. Inf. Model.* **54,** 470–480 (2014).

59. Park, S., Khalili-Araghi, F., Tajkhorshid, E. & Schulten, K. Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *J. Chem. Phys.* **119,** 3559–3566 (2003).

60. Laio, A. & Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports Prog. Phys.* **71,** 126601 (2008).

61. Bussi, G. & Laio, A. Using metadynamics to explore complex free-energy landscapes. *Nat. Rev. Phys.* **2,** 200–212 (2020).

62. Barducci, A., Bussi, G. & Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **100,** 020603 (2008).

63. Scheuermann, T. H., Stroud, D., Sleet, C. E., Bayeh, L., Shokri, C., Wang, H., Caldwell, C. G., Longgood, J., MacMillan, J. B., Bruick, R. K., Gardner, K. H. & Tambar, U. K. Isoform-Selective and Stereoselective Inhibition of Hypoxia Inducible Factor-2. *J. Med. Chem.* **58,** 5930–41 (2015).

64. Tiwary, P., Limongelli, V., Salvalaglio, M. & Parrinello, M. Kinetics of protein-ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E386–E391 (2015).

65. Casasnovas, R., Limongelli, V., Tiwary, P., Carloni, P. & Parrinello, M. Unbinding Kinetics of a p38 MAP Kinase Type II Inhibitor from Metadynamics Simulations. *J. Am. Chem. Soc.* **139,** 4780–4788 (2017).

66. Bernetti, M., Masetti, M., Recanatini, M., Amaro, R. E. & Cavalli, A. An Integrated Markov State Model and Path Metadynamics Approach to Characterize Drug Binding

Processes. *J. Chem. Theory Comput.* **15,** 5689–5702 (2019).

67. Bertazzo, M., Gobbo, D., Decherchi, S. & Cavalli, A. Machine Learning and Enhanced Sampling Simulations for Computing the Potential of Mean Force and Standard Binding Free Energy. *J. Chem. Theory Comput.* **17,** 5287–5300 (2021).

68. Hovan, L., Comitani, F. & Gervasio, F. L. Defining an Optimal Metric for the Path Collective Variables. *J. Chem. Theory Comput.* **15,** 25–32 (2019).

69. Warshel, A. & Levitt, M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **103,** 227–249 (1976).

70. Field, M. J., Bash, P. A. & Karplus, M. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comput. Chem.* **11,** 700–733 (1990).

71. Groenhof, G. Introduction to QM/MM simulations. *Methods Mol. Biol.* **924,** 43–66 (2013).

72. Marx, D. & Hutter, J. *Ab Initio Molecular Dynamics*. (Cambridge University Press, 2009). doi:10.1017/CBO9780511609633

73. von Lilienfeld, O. A., Tavernelli, I., Rothlisberger, U. & Sebastiani, D. Variational optimization of effective atom centered potentials for molecular properties. *J. Chem. Phys.* **122,** 014113 (2005).

74. Car & Parrinello. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **55,** 2471–2474 (1985).

75. Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H. & Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **41,** 429–452 (2012).

76. Doshi, U. & Hamelberg, D. Towards fast, rigorous and efficient conformational sampling of biomolecules: Advances in accelerated molecular dynamics. *Biochim. Biophys. Acta - Gen. Subj.* **1850,** 878–888 (2015).

77. Sabbadin, D. & Moro, S. Supervised molecular dynamics (SuMD) as a helpful tool to

depict GPCR-ligand recognition pathway in a nanosecond time scale. *J. Chem. Inf. Model.* **54,** 372–6 (2014).

78. Dickson, A. & Lotz, S. D. Multiple Ligand Unbinding Pathways and Ligand-Induced Destabilization Revealed by WExplore. *Biophys. J.* **112,** 620–629 (2017).

79. Zeller, F., Luitz, M. P., Bomblies, R. & Zacharias, M. Multiscale Simulation of Receptor–Drug Association Kinetics: Application to Neuraminidase Inhibitors. *J. Chem. Theory Comput.* **13,** 5097–5105 (2017).

80. Basciu, A., Malloci, G., Pietrucci, F., Bonvin, A. M. J. J. & Vargiu, A. V. Holo-like and Druggable Protein Conformations from Enhanced Sampling of Binding Pocket Volume and Shape. *J. Chem. Inf. Model.* **59,** 1515–1528 (2019).

81. Deb, I. & Frank, A. T. Accelerating Rare Dissociative Processes in Biomolecules Using Selectively Scaled MD Simulations. *J. Chem. Theory Comput.* **15,** 5817–5828 (2019).

82. Lu, H. & Schulten, K. Steered molecular dynamics simulations of force-induced protein domain unfolding. *Proteins Struct. Funct. Genet.* **35,** 453–463 (1999).

83. Milles, L. F., Schulten, K., Gaub, H. E. & Bernardi, R. C. Molecular mechanism of extreme mechanostability in a pathogen adhesin. *Science (80-.).* **359,** 1527–1533 (2018).

84. Hsin, J., Strümpfer, J., Lee, E. H. & Schulten, K. Molecular Origin of the Hierarchical Elasticity of Titin: Simulation, Experiment, and Theory. *Annu. Rev. Biophys.* **40,** 187–203 (2011).

85. Giorgino, T. & De Fabritiis, G. A high-throughput steered molecular dynamics study on the free energy profile of ion permeation through gramicidin A. *J. Chem. Theory Comput.* **7,** 1943–1950 (2011).

86. Nademi, Y., Tang, T. & Uludağ, H. Steered molecular dynamics simulations reveal a self-protecting configuration of nanoparticles during membrane penetration. *Nanoscale* **10,** 17671–17682 (2018).

87. Jorgensen, W. L. Pulled from a protein ' s embrace Closing in on evaders. *Nature* **466,** 42–43 (2010).

88. Colizzi, F., Perozzo, R., Scapozza, L., Recanatini, M. & Cavalli, A. Single-molecule pulling simulations can discern active from inactive enzyme inhibitors. *J. Am. Chem. Soc.* **132,** 7361–7371 (2010).

89. Jarzynski, C. Rare events and the convergence of exponentially averaged work values. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **73,** 1–10 (2006).

90. Laio, A. & Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports Prog. Phys.* **71,** 126601 (2008).

91. Bussi, G. & Branduardi, D. Free-Energy Calculations with Metadynamics: Theory and Practice. *Rev. Comput. Chem.* **28,** 1–49 (2015).

92. Bersten, D. C., Sullivan, A. E., Peet, D. J. & Whitelaw, M. L. bHLH-PAS proteins in cancer. *Nat. Rev. Cancer* **13,** 827–41 (2013).

93. Wu, D., Potluri, N., Lu, J., Kim, Y. & Rastinejad, F. Structural integration in hypoxia-inducible factors. *Nature* **524,** 303–309 (2015).

94. Kewley, R. J., Whitelaw, M. L. & Chapman-Smith, A. The mammalian basic helix-loop-helix/PAS family of transcriptional regulators. *Int. J. Biochem. Cell Biol.* **36,** 189–204 (2004).

95. Supuran, C. T. Carbonic anhydrases: novel therapeutic applications for inhibitors and activators. *Nat. Rev. Drug Discov.* **7,** 168–81 (2008).

96. Dewhirst, M. W., Cao, Y. & Moeller, B. Cycling hypoxia and free radicals regulate angiogenesis and radiotherapy response. *Nat. Rev. Cancer* **8,** 425–437 (2008).

97. McIntosh, B. E., Hogenesch, J. B. & Bradfield, C. A. Mammalian Per-Arnt-Sim proteins in environmental adaptation. *Annu. Rev. Physiol.* **72,** 625–645 (2009).

98. Denison, M. S., Soshilov, A. A., He, G., DeGroot, D. E. & Zhao, B. Exactly the same but different: promiscuity and diversity in the molecular mechanisms of action of the aryl hydrocarbon (dioxin) receptor. *Toxicol. Sci.* **124,** 1–22 (2011).

99. Giani Tagliabue, S., Faber, S. C., Motta, S., Denison, M. S. & Bonati, L. Modeling the binding of diverse ligands within the Ah receptor ligand binding domain. *Sci. Rep.* **9,**

1–14 (2019).

100. Scheuermann, T. H., Tomchick, D. R., Machius, M., Guo, Y., Bruick, R. K. & Gardner, K. H. Artificial ligand binding within the HIF2alpha PAS-B domain of the HIF2 transcription factor. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 450–5 (2009).

101. Rogers, J. L., Bayeh, L., Scheuermann, T. H., Longgood, J., Key, J., Naidoo, J., Melito, L., Shokri, C., Frantz, D. E., Bruick, R. K., Gardner, K. H., Macmillan, J. B. & Tambar, U. K. Development of Inhibitors of the PAS - B Domain of the HIF-2 α Transcription Factor. *J. Med. Chem.* **56,** 1739–1747 (2013).

102. Scheuermann, T. H., Li, Q., Ma, H., Key, J., Zhang, L., Chen, R., Garcia, J. A., Naidoo, J., Longgood, J., Frantz, D. E., Tambar, U. K., Gardner, K. H. & Bruick, R. K. Allosteric inhibition of hypoxia inducible factor-2 with small molecules. *Nat. Chem. Biol.* **9,** 271–276 (2013).

103. Key, J., Scheuermann, T. H., Anderson, P. C., Daggett, V. & Gardner, K. H. Principles of ligand binding within a completely buried cavity in HIF2alpha PAS-B. *J. Am. Chem. Soc.* **131,** 17647–54 (2009).

104. Wallace, E. M., Rizzi, J. P., Han, G., Wehn, P. M., Cao, Z., Du, X., Cheng, T., Czerwinski, R. M., Dixon, D. D., Goggin, B. S., Grina, J. A., Halfmann, M. M., Maddie, M. A., Olive, S. R., Schlachter, S. T., Tan, H., Wang, B., Wang, K., Xie, S., Xu, R., Yang, H. & Josey, J. A. A Small-Molecule Antagonist of HIF2 Is Efficacious in Preclinical Models of Renal Cell Carcinoma. *Cancer Res.* **76,** 5491–5500 (2016).

105. Chen, W., Hill, H., Christie, A., Kim, M. S., Holloman, E., Pavia-Jimenez, A., Homayoun, F., Ma, Y., Patel, N., Yell, P., Hao, G., Yousuf, Q., Joyce, A., Pedrosa, I., Geiger, H., Zhang, H., Chang, J., Gardner, K. H., Bruick, R. K., Reeves, C., Hwang, T. H., Courtney, K., Frenkel, E., Sun, X., Zojwalla, N., Wong, T., Rizzi, J. P., Wallace, E. M., Josey, J. A., Xie, Y., Xie, X.-J., Kapur, P., McKay, R. M. & Brugarolas, J. Targeting renal cell carcinoma with a HIF-2 antagonist. *Nature* **539,** 112–117 (2016).

106. Cho, H., Du, X., Rizzi, J. P., Liberzon, E., Chakraborty, A. A., Gao, W., Carvo, I., Signoretti, S., Bruick, R. K., Josey, J. A., Wallace, E. M. & Kaelin, W. G. On-target efficacy of a HIF-2α antagonist in preclinical kidney cancer models. *Nature* **539,** 107–

111 (2016).

107. Motta, S., Minici, C., Corrada, D., Bonati, L. & Pandini, A. Ligand-induced perturbation of the HIF-2 α : ARNT dimer dynamics. *PLoS Comput. Biol.* **14,** e1006021 (2018).

108. Callea, L., Bonati, L. & Motta, S. Metadynamics-Based Approaches for Modeling the Hypoxia-Inducible Factor 2α Ligand Binding Process. *J. Chem. Theory Comput.* **17,** 3841–3851 (2021).

109. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

110. Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R. & Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided. Mol. Des.* **27,** 221–234 (2013).

111. Bas, D. C., Rogers, D. M. & Jensen, J. H. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins* **73,** 765–83 (2008).

112. Schrödinger Release 2020-1: LigPrep, Schrödinger, LLC, New York, NY. (2020).

113. 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).

114. Jakalian, A., Jack, D. B. & Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **23,** 1623–1641 (2002).

115. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **25,** 247–260 (2006).

116. 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).

117. Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B. & Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism

from laptops to supercomputers. *SoftwareX* **1–2,** 19–25 (2015).

118. 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).

119. Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E. & Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11,** 3696–3713 (2015).

120. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., Dinola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81,** 3684–3690 (1984).

121. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126,** 014101 (2007).

122. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52,** 7182–7190 (1981).

123. 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).

124. Macromodel versione 10.3, Schrödinger LLC, New York. (2014).

125. Shivakumar, D., Williams, J., Wu, Y., Damm, W., Shelley, J. & Sherman, W. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J. Chem. Theory Comput.* **6,** 1509–1519 (2010).

126. Glide versione 6.2, Schrödinger LLC, New York. (2014).

127. Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., Sanschagrin, P. C. & Mainz, D. T. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **49,** 6177–6196 (2006).

128. Totrov, M. & Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.* **18,** 178–184 (2008).

129. 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).

130. Bonomi, M., Branduardi, D., Bussi, G., Camilloni, C., Provasi, D., Raiteri, P., Donadio, D., Marinelli, F., Pietrucci, F., Broglia, R. A. & Parrinello, M. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* **180,** 1961–1972 (2009).

131. Casasnovas, R., Limongelli, V., Tiwary, P., Carloni, P. & Parrinello, M. Unbinding Kinetics of a p38 MAP Kinase Type II Inhibitor from Metadynamics Simulations. *J. Am. Chem. Soc.* **139,** 4780–4788 (2017).

132. Tiwary, P., Limongelli, V., Salvalaglio, M. & Parrinello, M. Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proc. Natl. Acad. Sci.* **112,** E386–E391 (2015).

133. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* (1996). doi:10.1016/0263-7855(96)00018-5

134. Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W. F. & Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chemie Int. Ed.* **38,** 236–240 (1999).

135. Scheuermann, T. H., Stroud, D., Sleet, C. E., Bayeh, L., Shokri, C., Wang, H., Caldwell, C. G., Longgood, J., MacMillan, J. B., Bruick, R. K., Gardner, K. H. & Tambar, U. K. Isoform-Selective and Stereoselective Inhibition of Hypoxia Inducible Factor-2. *J. Med. Chem.* **58,** 5930–5941 (2015).

136. Koyasu, S., Kobayashi, M., Goto, Y., Hiraoka, M. & Harada, H. Regulatory mechanisms of hypoxia-inducible factor 1 activity: Two decades of knowledge. *Cancer Sci.* **109,** 560–571 (2018).

137. Wu, D., Su, X., Lu, J., Li, S., Hood, B. L., Vasile, S., Potluri, N., Diao, X., Kim, Y., Khorasanizadeh, S. & Rastinejad, F. Bidirectional modulation of HIF-2 activity through chemical ligands. *Nat. Chem. Biol.* **15,** 367–376 (2019).

138. Burendahl, S., Danciulescu, C. & Nilsson, L. Ligand unbinding from the estrogen

receptor: A computational study of pathways and ligand specificity. *Proteins Struct. Funct. Bioinforma.* **77,** 842–856 (2009).

139. Hu, X., Hu, S., Wang, J., Dong, Y., Zhang, L. & Dong, Y. Steered molecular dynamics for studying ligand unbinding of ecdysone receptor. *J. Biomol. Struct. Dyn.* **36,** 3819–3828 (2018).

140. Shen, J., Li, W., Liu, G., Tang, Y. & Jiang, H. Computational insights into the mechanism of ligand unbinding and selectivity of estrogen receptors. *J. Phys. Chem. B* **113,** 10436–10444 (2009).

141. Shirts, M. R., Mobley, D. L. & Brown, S. P. in *Free. Calc. Struct. drug Des.* (eds. Merz, K. M., Ringe, D. & Reynolds, C. H.) 61–86 (Cambridge University Press, 2010). doi:10.1017/CBO9780511730412.007

142. Bernetti, M., Masetti, M., Recanatini, M., Amaro, R. E. & Cavalli, A. An Integrated Markov State Model and Path Metadynamics Approach To Characterize Drug Binding Processes. *J. Chem. Theory Comput.* **15,** 5689–5702 (2019).

143. Fidelak, J., Juraszek, J., Branduardi, D., Bianciotto, M. & Gervasio, F. L. Free-energy-based methods for binding profile determination in a congeneric series of CDK2 inhibitors. *J. Phys. Chem. B* **114,** 9516–9524 (2010).

144. Capelli, R., Carloni, P. & Parrinello, M. Exhaustive Search of Ligand Binding Pathways via Volume-Based Metadynamics. *J. Phys. Chem. Lett.* **10,** 3495–3499 (2019).

145. Amaro, R. E., Baron, R. & McCammon, J. A. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J. Comput. Aided. Mol. Des.* **22,** 693–705 (2008).

146. Ellingson, S. R., Miao, Y., Baudry, J. & Smith, J. C. Multi-conformer ensemble docking to difficult protein targets. *J. Phys. Chem. B* **119,** 1026–34 (2015).

147. Motta, S. & Bonati, L. Modeling Binding with Large Conformational Changes: Key Points in Ensemble-Docking Approaches. *J. Chem. Inf. Model.* **57,** 1563–1578 (2017).

148. Teuvo Kojonen. The Self-organizing Map. *Proc. IEEE* **78,** 1464–1480 (1990).

149. Miljković, D. Brief Review of Self-Organizing Maps. 1061–1066 (2017).

150. Kohonen, T. Essentials of the self-organizing map. *Neural Networks* **37,** 52–65 (2013).

151. Pandini, A., Fraccalvieri, D. & Bonati, L. Artificial Neural Networks for Efficient Clustering of Conformational Ensembles and their Potential for Medicinal Chemistry. *Curr. Top. Med. Chem.* **13,** 642–651 (2013).

152. Oja, M., Kaski, S. & Kohonen, T. Bibliography of Self-Organizing Map (SOM) Papers : 1998-2001 Addendum. 1–156 (2002).

153. Kaski, S., Kangas, J. & Kohonen, T. Bibliography of self-organizing map (SOM) papers: 1981–1997. *Neural Comput. Surv.* **1,** 1–176 (1998).

154. Mallet, V., Nilges, M. & Bouvier, G. quicksom: Self-Organizing Maps on GPUs for clustering of molecular dynamics trajectories. *Bioinformatics* 0–2 (2020). doi:10.1093/bioinformatics/btaa925

155. Bouvier, G., Desdouits, N., Ferber, M., Blondel, A. & Nilges, M. An automatic tool to analyze and cluster macromolecular conformations based on self-organizing maps. *Bioinformatics* **31,** 1490–1492 (2015).

156. Fraccalvieri, D., Pandini, A., Stella, F. & Bonati, L. Conformational and functional analysis of molecular dynamics trajectories by self-organising maps. *BMC Bioinformatics* **12,** 158 (2011).

157. Fraccalvieri, D., Tiberti, M., Pandini, A., Bonati, L. & Papaleo, E. Functional annotation of the mesophilic-like character of mutants in a cold-adapted enzyme by self-organising map analysis of their molecular dynamics. *Mol. Biosyst.* **8,** 2680 (2012).

158. Mantsyzov, A. B., Bouvier, G., Evrard-Todeschi, N. & Bertho, G. Contact-based ligand-clustering approach for the identification of active compounds in virtual screening. *Adv. Appl. Bioinforma. Chem.* **5,** 61–79 (2012).

159. Harigua-Souiai, E., Cortes-Ciriano, I., Desdouits, N., Malliavin, T. E., Guizani, I., Nilges, M., Blondel, A. & Bouvier, G. Identification of binding sites and favorable ligand binding moieties by virtual screening and self-organizing map analysis. *BMC Bioinformatics* **16,** 11–15 (2015).

160. Joseph, A. P., Agarwal, G., Mahajan, S., Gelly, J. C., Swapna, L. S., Offmann, B., Cadet,

F., Bornot, A., Tyagi, M., Valadié, H., Schneider, B., Etchebest, C., Srinivasan, N. & de Brevern, A. G. A short survey on protein blocks. *Biophys. Rev.* **2,** 137–145 (2010).

161. De Brevern, A. G., Etchebest, C. & Hazout, S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins Struct. Funct. Genet.* **41,** 271–287 (2000).

162. Craveur, P., Joseph, A. P., Esque, J., Narwani, T. J., Noël, F., Shinada, N., Goguet, M., Leonard, S., Poulain, P., Bertrand, O., Faure, G., Rebehmed, J., Ghozlane, A., Swapna, L. S., Bhaskara, R. M., Barnoud, J., Téletchéa, S., Jallu, V., Cerny, J., Schneider, B., Etchebest, C., Srinivasan, N., Gelly, J. C. & de Brevern, A. G. Protein flexibility in the light of structural alphabets. *Front. Mol. Biosci.* **2,** 1–20 (2015).

163. Duclert-Savatier, N., Bouvier, G., Nilges, M. & Malliavin, T. E. Building Graphs to Describe Dynamics, Kinetics, and Energetics in the d -ALa: D -Lac Ligase VanA. *J. Chem. Inf. Model.* **56,** 1762–1775 (2016).

164. Motta, S., Pandini, A., Fornili, A. & Bonati, L. Reconstruction of ARNT PAS-B Unfolding Pathways by Steered Molecular Dynamics and Artificial Neural Networks. *J. Chem. Theory Comput.* **17,** 2080–2089 (2021).

165. Schultz, C. R., Geerts, D., Mooney, M., El-Khawaja, R., Koster, J. & Bachmann, A. S. Synergistic drug combination GC7/DFMO suppresses hypusine/spermidine-dependent eIF5A activation and induces apoptotic cell death in neuroblastoma. *Biochem. J.* **475,** 531–545 (2018).

166. D'Agostino, M., Motta, S., Romagnoli, A., Orlando, P., Tiano, L., La Teana, A. & Di Marino, D. Insights Into the Binding Mechanism of GC7 to Deoxyhypusine Synthase in Sulfolobus solfataricus: A Thermophilic Model for the Design of New Hypusination Inhibitors. *Front. Chem.* **8,** 1–14 (2020).

167. Pons, P. & Latapy, M. in *Iscis* 284–293 (2005). doi:10.1007/11569596_31

168. Spedicato, G. A. Discrete Time Markov Chains with R. *R J.* **9,** 84–104 (2017).

169. Metzner, P., Schütte, C. & Vanden-Eijnden, E. Transition Path Theory for Markov Jump Processes. *Multiscale Model. Simul.* **7,** 1192–1219 (2009).

170. Wehrens, R. & Kruisselbrink, J. Flexible self-organizing maps in kohonen 3.0. *J. Stat. Softw.* **87,** (2018).

171. Wehrens, R. Self- and Super-Organizing Maps in R: The Kohonen Package. *JSS J. Stat. Softw.* **21,** (2007).

172. Csárdi, G. & Tamás, N. The Igraph Software Package for Complex Network Research. *InterJournal* **Complex Sy,** 1695 (2006).

173. Motta, S., Callea, L., Bonati, L. & Pandini, A. PathDetect-SOM: A Neural Network Approach for the Identification of Pathways in Ligand Binding Simulations. *J. Chem. Theory Comput.* Accepted with major revision.

174. Freeman, L. C. Centrality in social networks conceptual clarification. *Soc. Networks* **1,** (1978).

175. Abraham, M. J., Murtola, T., Schulz, R., Smith, J. C., Hess, B. & Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. **2,** 19–25 (2015).

176. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9,** 3084–3095 (2013).

177. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **32,** 2319–2327 (2011).

178. Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A. & Caves, L. S. D. Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics* **22,** 2695–2696 (2006).

179. McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L. P., Lane, T. J. & Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109,** 1528–1532 (2015).

180. Husic, B. E. & Pande, V. S. Markov State Models: From an Art to a Science. *J. Am.*

Chem. Soc. **140,** 2386–2396 (2018).

181. 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).

182. Kieninger, S., Donati, L. & Keller, B. G. Dynamical reweighting methods for Markov models. *Curr. Opin. Struct. Biol.* **61,** 124–131 (2020).

183. Kairys, V., Baranauskiene, L., Kazlauskiene, M., Matulis, D. & Kazlauskas, E. Binding affinity in drug design: experimental and computational techniques. *Expert Opin. Drug Discov.* **14,** 755–768 (2019).

184. Copeland, R. A. The drug–target residence time model: a 10-year retrospective. **15,** 87–95 (2016).

185. Tiwary, P. & Parrinello, M. A time-independent free energy estimator for metadynamics. *J. Phys. Chem. B* **119,** 736–742 (2015).

186. Limongelli, V. Ligand binding free energy and kinetics calculation in 2020. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **10,** 1–32 (2020).

187. Chodera, J. D. & Noé, F. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **25,** 135–144 (2014).

188. Olsson, S. & Noé, F. Dynamic graphical models of molecular kinetics. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 15001–15006 (2019).

189. Plattner, N. & Noé, F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat. Commun.* **6,** (2015).

190. Capelli, R., Lyu, W., Bolnykh, V., Meloni, S., Olsen, J. M. H., Rothlisberger, U., Parrinello, M. & Carloni, P. Accuracy of Molecular Simulation-Based Predictions of koffValues: A Metadynamics Study. *J. Phys. Chem. Lett.* **11,** 6373–6381 (2020).

191. 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).

192. Chiariello, M. G., Bolnykh, V., Ippoliti, E., Meloni, S., Olsen, J. M. H., Beck, T.,

Rothlisberger, U., Fahlke, C. & Carloni, P. Molecular Basis of CLC Antiporter Inhibition by Fluoride. *J. Am. Chem. Soc.* **142,** 7254–7258 (2020).

193. Rizzi, A., Carloni, P. & Parrinello, M. Targeted Free Energy Perturbation Revisited: Accurate Free Energies from Mapped Reference Potentials. *J. Phys. Chem. Lett.* **12,** 9449–9454 (2021).

194. Bolnykh, V., Olsen, J. M. H., Meloni, S., Bircher, M. P., Ippoliti, E., Carloni, P. & Rothlisberger, U. Extreme Scalability of DFT-Based QM/MM MD Simulations Using MiMiC. *J. Chem. Theory Comput.* **15,** 5601–5613 (2019).

195. Chiariello, M. G., Alfonso-Prieto, M., Ippoliti, E., Fahlke, C. & Carloni, P. Mechanisms Underlying Proton Release in CLC-type F-/H+Antiporters. *J. Phys. Chem. Lett.* **12,** 4415–4420 (2021).

196. Braka, A., Garnier, N., Bonnet, P. & Aci-Sèche, S. Residence Time Prediction of Type 1 and 2 Kinase Inhibitors from Unbinding Simulations. *J. Chem. Inf. Model.* **60,** 342–348 (2020).

197. Huang, Y. ming M. Multiscale computational study of ligand binding pathways: Case of p38 MAP kinase and its inhibitors. *Biophys. J.* **120,** 3881–3892 (2021).

198. Zhang, D., Huang, S., Mei, H., Kevin, M., Shi, T. & Chen, L. Protein–ligand interaction fingerprints for accurate prediction of dissociation rates of p38 MAPK Type II inhibitors. *Integr. Biol.* **11,** 53–60 (2019).

199. Nunes-Alves, A., Ormersbach, F. & Wade, R. C. Prediction of the Drug-Target Binding Kinetics for Flexible Proteins by Comparative Binding Energy Analysis. *J. Chem. Inf. Model.* **61,** 3708–3721 (2021).

200. Olsen, J. M. H., Bolnykh, V., Meloni, S., Ippoliti, E., Bircher, M. P., Carloni, P. & Rothlisberger, U. MiMiC: A Novel Framework for Multiscale Modeling in Computational Chemistry. *J. Chem. Theory Comput.* **15,** 3810–3823 (2019).

201. Bolnykh, V., Olsen, J. M. H., Meloni, S., Bircher, M. P., Ippoliti, E., Carloni, P. & Rothlisberger, U. MiMiC: Multiscale Modeling in Computational Chemistry. *Front. Mol. Biosci.* **7,** 1–4 (2020).

202. Bolnykh, V., Rossetti, G., Rothlisberger, U. & Carloni, P. Expanding the boundaries of ligand–target modeling by exascale calculations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **11,** 1–17 (2021).

203. Capelli, R., Lyu, W., Bolnykh, V., Meloni, S., Olsen, J. M. H., Rothlisberger, U., Parrinello, M. & Carloni, P. Accuracy of Molecular Simulation-Based Predictions of k off Values: A Metadynamics Study. *J. Phys. Chem. Lett.* **11,** 6373–6381 (2020).

204. CPMD, Copyright IBM Corp 1990−2018, Copyright MPI für Festkörperforschung, Stuttgart, 1997−2001. http://www.cpmd.org/.

205. Asih, P. R., Prikas, E., Stefanoska, K., Tan, A. R. P., Ahel, H. I. & Ittner, A. Functions of p38 MAP Kinases in the Central Nervous System. *Front. Mol. Neurosci.* **13,** 1–27 (2020).

206. Machado, T. R., Machado, T. R. & Pascutti, P. G. The p38 MAPK Inhibitors and Their Role in Inflammatory Diseases. *ChemistrySelect* **6,** 5729–5742 (2021).

207. Pargellis, C., Tong, L., Churchill, L., Cirillo, P. F., Gilmore, T., Graham, A. G., Grob, P. M., Hickey, E. R., Moss, N., Pav, S. & Regan, J. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol.* **9,** 268–272 (2002).

208. Dataset at <https://github.com/deGrootLab/pmx/tree/master/protLig_benchmark/p38/transformations_gaff2>.

209. Gapsys, V., Pérez-Benito, L., Aldeghi, M., Seeliger, D., Van Vlijmen, H., Tresadern, G. & De Groot, B. L. Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chem. Sci.* **11,** 1140–1152 (2020).

210. JURECA Compute cluster. at <http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JURECA/Configuration/ Configuration_node.html>

211. JUWELS Compute cluster. at <http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUWELS/Configuration/ Configuration_node.html>

212. Troullier, N. & Martins, J. L. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B* **43,** 1993–2006 (1991).

213. Becke, A. D. Density-functional exchange-energy approximation with correct

asymptotic behavior. *Phys. Rev. A* **38,** 3098–3100 (1988).

214. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37,** 785–789 (1988).

215. Best, R. B. & Hummer, G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J. Phys. Chem. B* **113,** 9004–15 (2009).

216. Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O. & Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinforma.* **78,** 1950–1958 (2010).

217. Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **52,** 255–268 (1984).

218. Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **31,** 1695–1697 (1985).

219. Adasme, M. F., Linnemann, K. L., Bolz, S. N., Kaiser, F., Salentin, S., Haupt, V. J. & Schroeder, M. PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. *Nucleic Acids Res.* **49,** W530–W534 (2021).

220. Petersson, G. A., Bennett, A., Tensfeldt, T. G., Al-Laham, M. A., Shirley, W. A. & Mantzaris, J. A complete basis set model chemistry. I. The total energies of closed-shell atoms and hydrides of the first-row elements. *J. Chem. Phys.* **89,** 2193–2218 (1988).

221. Petersson, G. A. & Al-Laham, M. A. A complete basis set model chemistry. II. Open-shell systems and the total energies of the first-row atoms. *J. Chem. Phys.* **94,** 6081–6090 (1991).

222. Gaussian 09, Revision A.02, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci.

223. Carloni, P., Andreoni, W., Hutter, J., Curioni, A., Giannozzi, P. & Parrinello, M. Structure and bonding in cisplatin and other Pt(II) complexes. *Chem. Phys. Lett.* **234,** 50–56 (1995).

224. Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J.*

Chem. Theory Comput. **4,** 116–122 (2008).

225. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126,** 014101 (2007).

“None of us got to where we are alone.
Whether the assistance we received was obvious or subtle,
acknowledging someone’s help is a big part of understanding
the importance of saying thank you.”

Harvey Mackay

ACKNOWLEDGEMENT

At the end of this work, I would like to dedicate a few sentences to thank those who have contributed to the success of this project and those who have supported me.

My first thanks is to my supervisor, Laura Bonati, who has guided and supported me not only during these years of my PhD but also since my Bachelor degree. Thank you, Laura, for your teachings, for your patience, for our discussions and for all the times you have encouraged me to do my best.

A big thanks to Dr. Stefano Motta, whom I prefer to call Ste, for transferring his knowledge to me with perseverance, determination, and patience. Thank you for helping and supporting me not only in my work but also in the difficult moments of this last year. I consider you a friend. I would also like to thank your wonderful wife, Raffaella, for always listening and understanding me; and staying within the family, thanks also to your beautiful babies, Gloria and Luna, who have always been able to put a smile on my face.

I wouldn't have got to this point if I didn’t have two special parents. Thank you mum for your constant support and immeasurable love. Thank you, above all, because together we have got up again and we have found a new stability. Thank you, Dad, because, even though you have passed away, every time I close my eyes you smiling at me and I can feel you close to my heart.

Thanks to you Peppe, my future husband, for always finding the perfect word to bring me up. Thank you for having supported and tolerated me, sharing my joys and pains. You are special.

Thanks also to my Sesina and Salvatore, essential presences in our lives.

To my dearest friends, near and far, thank you for always being at my side.

Finally, a thank you to Prof. Paolo Carloni and his group with whom I collaborated during the last year of my PhD. Thank you for having welcomed me "virtually". The pandemic impeded me to be in Jūlich with you, but I learned a lot from our discussions.