SCUOLA DI DOTTORATO

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Earth and Environmental Sciences

PhD Program in Chemical, Geological and Environmental Sciences

Cycle XXX

Curriculum in Chemical Sciences

PROTEIN DYNAMICS SIMULATIONS TO INVESTIGATE BIOMOLECULAR INTERACTIONS

Motta Stefano

Matr. 728520

Tutor: Prof.ssa Laura Bonati

Coordinator: Prof.ssa Maria Luce Frezzotti

Academic Year: 2016/2017

CONTENTS

ABBREVIATIONS

Proteins and Molecules

AChBP	Acetylcholine Binding Protein
nAChRs	Nicotinic Acetylcholine Receptors
Allose BP	Allose Binding Protein
EPJ	Hepes
EPE	Epibatidine
LOB	Lobeline
COC	Cocaine
MLK	Methyllycaconitine
HIF-2α	Hypoxia Inducible Factor 2α
ARNT	Aryl Hydrocarbon Receptor Nuclear Translocator
AhR	Aryl Hydrocarbon Receptor
AhRR	Aryl Hydrocarbon Receptor Repressor
CLOCK	Circadian Locomotor Output Cycles Kaput
NPAS	Neuronal PAS Domain Protein
SIM	Single Minded
0X3	N-(3-chloro-5-fluorophenyl)-4-nitro-2,1,3-benzoxadiazol-5-amine
A2aR	Adenosine A2A Receptor
GPCR	G-Protein Coupled Receptor
POPC	1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine

Computational Techniques

MM	Molecular Mechanics
QM	Quantum Mechanics
MD	Molecular Dynamics

RESP	Restricted Electrostatic Potential
GAFF	Generalized Amber Force Field
PBC	Periodic Boundary Condition
PME	Particle Mesh Ewald
aMD	Accelerated Molecular Dynamics
cMD	Conventional Molecular Dynamics
MRC	Multiple Receptor Conformations
Glide SP	Standard Precision Glide
Glide XP	Extra Precision Glide
KIC	Kinematic Closure
NGK	Next Generation Kinematic closure
SOM	Self Organizing Map
MM-GBSA	Molecular Mechanics Generalized Born Surface Area
GB	Generalized Born
SA	Surface Area
LCPO	Linear Combination of Pairwise Overlap
PC	Principal Components
DCCM	Distance Cross Correlation Matrix
CG	Coarse-Grained
CG-MetaD	Coarse-Grained Metadynamics
US	Umbrella Sampling
EN	Elastic Network
ELNEDIN	Elastic Network Dynamic
WT	Well Tempered
MW	Multiple Walkers

Experimental Techniques

NMR	Nuclear Magnetic Resonance
co-IP	Co-immunoprecipitation
BRET	Bioluminescence Resonance Energy Transfer
FRET	Fluorescence Resonance Energy Transfer
EM	Electron Microscopy

General and Miscellaneous

PDB	Protein Data Bank
PES	Potential Energy Surface
FES	Free Energy Surface
PMF	Potential of Mean Force
RMSD	Root Mean Square Deviation
dRMSD	Distance Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
IF	Induced Fit
CS	Conformational Selection
CV	Collective Variable
bHLH	Basic Helix–Loop–Helix
PAS	PER/ARNT/SIM
TAD	Transactivation Domain
ODD	Oxygen Dependent Degradation Domain
HREs	Hypoxia Response Elements
VEGF	Vascular Endothelial Growth Factor
EMT	Epithelial–Mesenchymal Transition
VHL	Von Hippel-Lindau
TM	Transmembrane helical segment
H8	Intracellular Amphipathic helix 8
VMD	Visual Molecular Dynamics

"The purpose of computing
is insight, not numbers."

Hamming, Wesley Richard -
Numerical Methods for Scientists and Engineers (1962)

CHAPTER 1

INTRODUCTION

1.1 Protein dynamics in biomolecular interactions and functions

Biology largely runs on the amazing tricks proteins can perform, but looking closely to them, proteins are simply molecules obeying the laws of physics and chemistry. We can think of them as machines, but there is no ghost inside.[1] Proteins are completely *soulless* objects, that fluctuate between different conformations.[2] Biology, via evolution, has indeed selected for highly useful structural fluctuations but, to understand these highly evolved functions, their spontaneous

"*jigglings and wigglings*" should be considered.[3] Structural determination of biomolecules is indispensable to understand proteins in their biological roles, and X-ray crystallography is currently the most important experimental technique to elucidate protein structures. A major drawback of crystal structures, however, is the representation of highly flexible macromolecules as a single static conformation,[4] while is the dynamical behaviour of proteins that allows them to act as signalling molecules, transporters, catalysts, sensors, and mechanical effectors.[5] Proteins are not isolated systems: they interact dynamically with hormones, drugs, substrates and one another. The ability of proteins to interact with other molecules is exactly what opens a wide range of chances to exert and regulate biological mechanisms in nature. Biological processes are indeed carried out through binding. Transmitting a signal, and forming an active molecular species, are examples of the diverse outcomes of binding events.[6]

Considerable research efforts have been addressed to investigate the nature of these mechanisms for both ligand-protein and protein-protein binding. The simple *lock and key* model has been widely used for the description of several binding mechanisms in which proteins do not undergo significant conformational changes, while the *induced fit* (IF) and *conformational selection* (CS) models have been introduced to handle binding to flexible targets.[7,8] The IF model relies on the hypothesis that the interaction between a protein and its binding partner induces a conformational change in the protein.[9] The CS theory describes molecular recognition as a process in which the binding partner selects the most complementary receptor conformation from an ensemble of pre-existing metastable states, which in turn shifts the dynamic population equilibrium towards the conformation adopted in the bound state.[10,11] Thus these two models can be identified by the temporal ordering of the binding step and the conformational change along the pathway: according to IF the conformational change occurs after the binding step while CS expects

that conformational change occurs prior to the binding step. Each of the above mechanisms play important roles in molecular recognition and has its own range of applicability for specific systems and under certain conditions. The extent of protein conformational change and the kind of interactions involved are both key element that can influence the model involved in binding.[12] On the other hand, the two mechanisms may coexist in the same process. In fact, in several cases, it has been shown that a protein scaffold close to the bound conformation is chosen through conformational selection and subsequently further changes occur to optimize the intermolecular interactions by induced fit.[13,14]

Upon the binding event, the protein conformation (and dynamics) can be modified even in spatial regions far from the binding site. These cases are usually referred to as *allostery*.[15–17] Allosteric communication is the basis for signalling within the cell and signals can traverse single-chain proteins, large multi-molecular complexes, and pathways, through pathway cross- talk, traveling across the entire cellular network.[18] Unravelling the complex and mazy network of biomolecular interactions that regulate protein function would mean understanding the complex machinery of cell. The specific function of a protein is indeed determined by the extent to which a macromolecule populates its active conformation, and binding of other molecules can modify this equilibrium (Figure 1.1).[19,20] It is important to understand how this happens to improve the development of novel *allosteric* drugs that modulate protein functions. Unlike *orthosteric* drugs, that binds into an active site and block it, allosteric drugs bind elsewhere on the protein and alter the population of the protein conformations. While orthosteric drugs shut off native protein function, allosteric ones can modulate it, either as agonists or antagonists. Anyway, the discovery of allosteric drugs is a challenging task. The chemical difference between an agonist and an antagonist can be of simply one atom and unlike orthosteric drugs, for which a key determinant of drug outcome is high affinity, in the

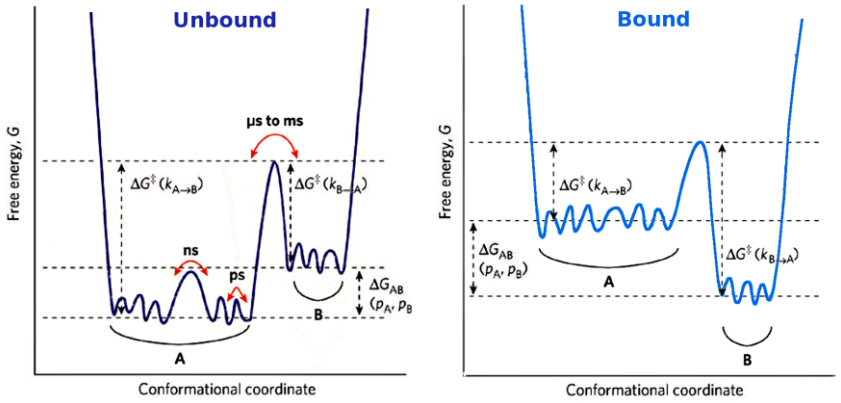

Figure 1.1: Free energy landscapes of illustrative unbound (left) and bound (right) states. The binding of a molecule such as another protein, ligands, or DNA can modify the free energy landscape of the system, altering the equilibrium between different states and causing an alteration in protein function. On the left, the timescales of transitions are indicated over the corresponding barriers (Image modified from ref [4]).

case of allosteric modulators the extent of the stabilization of the active (or inactive) conformation (allosteric efficacy) can be pivotal in specifying the drug action.[21]

The tools available to scientists to study processes happening on microscopic length-scale have increased in the last decades. Structural information can be obtained from X-ray diffraction, nuclear magnetic resonance (NMR) or electron microscopy. The *form follows function* principle, indeed, implies that a structural determination of proteins (form) is essential to understand proteins in their biological roles (function).[22] It is however difficult to obtain dynamical information about proteins using these methods. Some of them only provide static (or average) information, while others are limited in their spatial and temporal resolution[4] (Figure 1.2). An appealing alternative to experiments is the *in-silico* study of protein structure, dynamics and binding. This is usually accomplished by approaches based on molecular mechanics (MM), that allow to study biological systems from thousands up to millions of atoms at atomistic resolution and to explore their free-energy landscape. While electronic structure

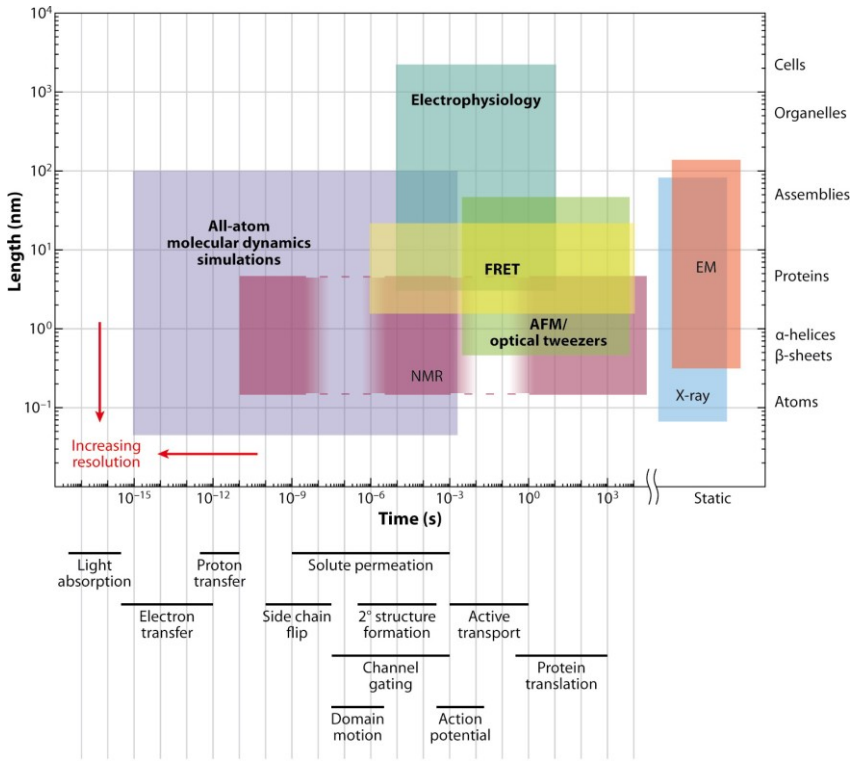

Figure 1.2: Spatiotemporal resolution of various biophysical techniques (Image from ref [5]).

calculations of entire proteins with quantum mechanical (QM), semi-empirical or DFT, methods are possible nowadays and allow the treatment of large systems of up to 150000 atoms, a very limited exploration of the phase space is allowed with these methods. Investigations of QM phenomena in realistic biological systems, solvated in biologically representative environments, can only be achieved using multiscale hybrid approaches, in which only a portion of the system is treated at the QM level while the rest of the system is represented with an empirical force field (QM/MM approaches).[23]

A first class of methods based on MM potentials that enable the study of ligand-protein and protein-protein interactions, even in the absence of a 3D structure of the complex, are docking methods. Since its beginning,[24] docking has evolved from the original description of

partners as rigid bodies, to modern approaches that allow to include flexibilities on both sides.[25,26] Protein-ligand docking, in particular, has gained an increasing audience through years for its relevance in structure-based drug-discovery studies. Fast exploration of different binding modes and the use of (semi)empirical scoring functions able to identify the best binding mode and to rank ligands by affinity are the strengths of these approaches. Given the speed of calculation (often from seconds to minutes on a simple workstation) these methods find their primary application in virtual screening campaigns, where thousands of ligands are screened versus a rigid receptor conformation to identify possible lead compounds.[27,28] However, the main limitation of these approaches is the treatment of the protein receptor as a rigid body. The inclusion of the protein flexibility, indeed, implies an explosion of degrees of freedom to be sampled by docking and thus it would require an amount of computation often out of reach of modern computers. Several methods that allow to include protein flexibility to some degree has been introduced, ranging from soft-docking approaches,[29] to docking in experimental *multiple receptor conformations* (MRC),[30] to methods that treat side-chain flexibility[31–33] or even backbone flexibility[34] on the fly. Docking anyway is not only limited to protein-ligand interactions. A number of docking approaches have been developed during the years able to model protein-protein interactions, ranging from rigid-body or soft rigid-body approximations to "high-resolution" methods that consider protein flexibility.[35] Flexible regions may be limited to local interface regions, or global search in the high-dimensional conformational space may be performed. In the latter case, Monte Carlo[36–38] or Molecular Dynamics[39,40] methods have been widely used, or normal modes[41–44] have been proposed to describe the conformational space. Also for protein-protein docking, partially flexible docking may be performed with an ensemble of protein conformations.

Despite the efforts made in the last years, docking methods suffer of several weaknesses that limit their applicability, and one of the most important problems regard the treatment of conformational flexibility, mainly in the cases of large conformational changes of the protein(s).[45–47] Moreover, docking procedures alone do not provide mechanistic information, and cannot explain the propagation of the binding effects. To overcome these limitations docking is increasingly coupled to (or even replaced by) molecular dynamics (MD) simulations.[48] In MD simulations the motion of atoms is studied integrating the Newton's law of motion. The method has been widely used since the first pioneering studies by Karplus and McCammon[49] and by Warshel and Levitt,[50] playing a crucial role in the computational studies of biological mechanisms. MD can be used, for example, to generate protein conformational ensembles to be used in MRC docking approaches.[51] This choice, that borrows from the CS model, introduces receptor flexibility before docking. An alternative approach is to perform a post-docking MD refinement, to model the IF effects.[52] In these methods docking is still used to place the ligand in the binding site and protein flexibility is introduced before or after the docking calculation.

An alternative solution to study binding events is to simulate the whole binding process with MD. This approach consists in simulating the unbound systems for long enough to let the partners diffuse into the solvation waters until they find the way to form a stable complex. For slow binding processes this may exceed today-affordable simulation time (up to hundreds of microseconds and milliseconds). With the increasing power of computational resources and the advent of GPU acceleration, this kind of simulations have becoming possible and various MD simulations of the whole binding process have been reported both for protein-ligand[53–55] and protein-protein assembly.[56,57] In principle these simulations allow to calculate the free energy difference between bound and unbound states and the kinetic

constant of binding, and can also reveal important mesostates formed during the binding process.

MD simulations also allow to investigate the effect of binding on the entire system, elucidating the mechanism mediated by a biomolecular complex at the atomic level. Because MD simulations produce large amounts of data, and because interdependent motions can be subtle, protein coupled motions can be difficult to pick out. Several methods have been developed specifically to find residues that are allosterically coupled,[58–60] using metrics such as mutual information[61,62], interaction energies[63], or correlated motions[64–67]. In this last group, protein residues are depicted as graph nodes connected by edges with lengths that are inversely proportional to the correlation between their motions. Pathways between source and sink nodes can be identified, therefore connecting the source of a perturbation with the effect.[68] With these methods alterations of internal protein communication can be detected and the origin of perturbation identified, thus facilitating the rational design of allosteric drugs.

One of the main limitation of MD is its ability of adequately sampling motions happening on long timescales. Binding events and large conformational changes of proteins can often exceed the millisecond and second timescale, thus overcoming the actual computer power. Every conformational change is associated to a free-energy barrier that determines the rate of the transition (thus the timescale of the process). The higher is the barrier to overcome, the lower is the probability to observe the transition (Figure 1.1). Such limitations can lead to inadequate sampling of conformational states, which in turn limits the ability to analyse and reveal functional properties of the systems being examined. All relevant states of a system must be reached in simulations in order for its dynamics and function to be meaningfully characterized. To this aim, several enhanced-sampling techniques have been developed able to speed up

the sampling of slow motions of the system in MD simulations[69,70]. A first approach consists in representing the system with a coarse-grained model in which a group of atoms is descripted by a single particle.[71] This allows to reach simulation lengths up to milliseconds. Other methods allow to accelerate barrier-crossing while maintaining an atomistic description of the system. A simple way to obtain this speed-up is raising the simulation temperature, thus incrementing the energy available to the system. *Simulated annealing*[72] and *replica exchange*[73] methods are based on this strategy: the first consists of simulating at high temperature and then gradually cooling the system to trap it into its global minimum, while in the second different replicas of the system are simulated at different temperatures that are exchanged according to a Monte Carlo algorithm. Other methods make use of a bias potential that guide the system in the sampling of high energy states (Figure 1.3). In this category fall *steered MD*,

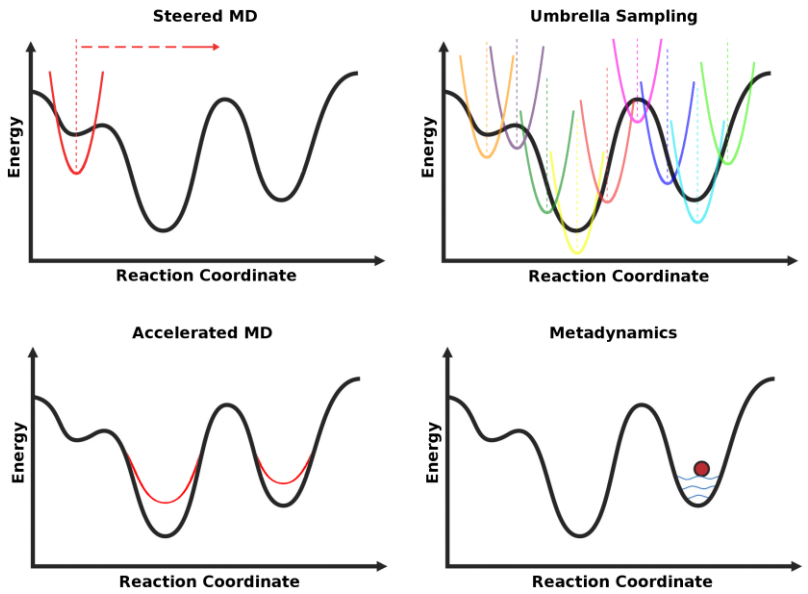

Figure 1.3: Schematic explanation of some biased enhanced-sampling methods. In steered MD the bias is moved along a specific reaction coordinate; umbrella sampling performs different replicas that sample overlapping windows; accelerated MD raises the energy of low-energy states; metadynamics adds an history-dependent bias that discourage re-sampling. (Image modified from ref [69])

umbrella sampling, *accelerated MD* and *metadynamics*. Steered MD employs a pulling force to bias the conformational change along a reaction coordinate.[74] In umbrella sampling separate simulations are run, each biased with a harmonic potential to a specific range of the reaction coordinate (window).[75] Accelerated MD adds a bias potential to raise low-energy states, reducing the energy required for the transition.[76] Finally, in metadynamics an history-dependent bias is added to the system in the form of Gaussian potentials to fill up the valley of the energy landscape.[77] All these techniques allow to compute the underlying free energy landscape of the system analysed, providing important information about thermodynamics and kinetics of the process of interest.

1.2 Outline of the thesis

The work presented in this PhD thesis deals with various aspects of biomolecular interactions that involve the dynamical behaviour of proteins. According to the specific problem addressed, different computational approaches, based on Molecular Dynamics simulations, have been applied.

In Chapter 2 (Theory of Molecular Dynamics Simulations), a general introduction to molecular simulations is provided, along with some hints of the use of statistical mechanics to derive average properties of the systems. Moreover, a more detailed description of some enhanced-sampling methods used in the PhD project is reported. Details of the specific methods selected for studying each of the investigated processes are instead presented and discussed in the related chapter.

In Chapter 3 (Modelling Binding with Large Conformational Changes: Key Points in Ensemble-Docking Approaches), ligand-binding processes associated with large conformational changes of the

receptor are discussed. The ensemble-docking strategy is suggested for treating such processes and the critical choices for effective prediction of the structure of ligand-protein complexes with this approach are discussed. Two study cases in which binding involves different conformational changes of the receptor and different ligand-protein interactions were selected: the acetylcholine binding protein (AChBP) mechanism, a prototypical example of non-specific binding of multiple diverse ligands at the same site;[78] and the allose binding protein (Allose BP) ligand binding, driven by the hinge-bending motion of two domains.[79,80] The availability of the X-ray structures of both the apo and holo forms made the above systems ideal test cases for our study. The protein structural ensembles for docking were obtained by MD simulations of the apo proteins. Given the high energy barriers separating protein bound and unbound conformational states, the accelerated MD (aMD) method was used to speed up the sampling. The optimal acceleration parameters of aMD simulations, the most effective cluster analysis procedure to extract relevant receptor conformations, and the best docking and refinement strategies are discussed for the two study cases.[81]

In Chapter 4 (Molecular Dynamics of HIF-2α:ARNT Ligand-Induced Inhibition), the problem of protein-protein interactions and their inhibition by ligands is tackled. The study case was the heterodimer of the hypoxia inducible factor 2α (HIF-2α) with the aryl hydrocarbon receptor nuclear translocator (ARNT). HIF-2α is a transcription factor that mediates the physiological responses to hypoxia. Given the importance of its function in tumour cells, where the oxygen concentration is often very low, disrupting protein-protein interactions of this heterodimer could be a potential therapeutic strategy, but directly interfering with dimer formation can be troublesome because of the difficulty to design drugs that bind to protein interfaces. However, ligands that bind internal protein cavities can indirectly perturb the interfaces reducing dimers stability, and several candidate inhibitors have been designed in last years.[82–84]

Albeit protein crystallography had offered a detailed static picture of a HIF-2α dimer bound to these inhibitors,[85] it is not able to describe either the perturbation caused by binding or the molecular mechanism of dimer destabilization. We exploited MD simulations to characterize the dynamical and energetic properties of the dimerization interfaces in both the unbound and inhibitor-bound systems to shed light on the ligand-induced perturbations. Moreover, we analysed correlated motions and inter-domain communication paths in the two systems to reveal the mechanism of ligand allosteric inhibition at atomic detail. These findings will guide toward the design of improved dimerization inhibitors.[86]

In Chapter 5 (Investigation of Adenosine A2A Receptor Dimerization Through Coarse-Grained Metadynamics), the study of the binding event of two membrane proteins with the use of MD enhanced-sampling methods is presented. The protein object of study was the adenosine A2A receptor (A2aR), a membrane protein belonging to the G-protein coupled receptor (GPCR) superfamily. This protein family is of incredible pharmaceutical importance due to the role of cell sensors played by its members. In the last years the discovery of supramolecular organization of GPCRs within the cell surface has opened the question of possible functional roles of dimers and oligomers of these proteins. The binding and unbinding events between two proteins embedded into the cell membrane is a slow process which timescale is beyond the possibility of conventional MD simulations. To speed up calculations we used a combination of coarse-grained representation of the system and metadynamics simulations. These studies allowed us to characterize the possible protein-protein interfaces of an A2aR dimer, and the associated binding free-energy landscape.

Finally, in Chapter 6 (Conclusions) some concluding remarks are reported to summarize the main findings deduced from the whole PhD

project, regarding the study of protein dynamics in biomolecular interactions for the comprehension of biological mechanisms.

"...if we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms."

Richard Feynman,
The Feynman Lectures on Physics (1964)

CHAPTER 2

THEORY OF MOLECULAR DYNAMICS SIMULATIONS

Proteins exist as a dynamical equilibrium among different conformations and experiments can only provide macroscopic observable associated to the microscopic behaviour of protein conformational ensembles.[2] Computer simulations can investigate the dynamics of proteins at an atomistic level, thus leading to the comprehension of biological mechanisms.[87,88] A complete description of the physics of the system can only be obtained if quantum mechanical approaches are used for computing the energy and the

forces that govern the system. However, these methods are limited in the number of atoms that can be taken into account, due to the high computational cost. Valid alternatives are provided by classical mechanics methods that approximate atoms as rigid spheres governed by the Newton laws and allow the study of large systems composed of thousands of atoms.

In this chapter, the basics principle of molecular dynamics (MD) simulations are presented, including: the molecular mechanics (MM) approximation, computational techniques for energy minimization, theory and setup of MD simulations, basic notions of statistical mechanics, and some advanced methods for a rapid exploration of the free energy landscape of the systems that have been used for the PhD project.

2.1 Molecular Mechanics

In the MM methods,[89] classical type models are used to predict the energy of a molecule as a function of its conformation. Atoms are approximated as particles with a given mass, radius and charge, but electrons are not explicitly considered. The functional form of the potential energy can be expressed as a sum of different terms:

$$U = U_{str} + U_{bend} + U_{tors} + U_{el} + U_{vdW}$$

each of which is a function of the coordinates of the system. The first three terms represent the covalent part of the energy (stretching, bending and torsional potentials, respectively), while the last two describe the non-bonded interactions between non-bonded atoms (electrostatic and van der Waals).

To simulate the existence of a bond between two atoms, bonded interactions are treated as springs subjected to the Hooke law. Thus,

the potential energy associated to the vibration of a bond is described by a harmonic potential:

$$U_{str_{(r)}} = \frac{1}{2}k\left(r - r_{eq}\right)^2$$

which has a minimum at the reference bond length r_{eq} and the force constant k regulate the stretching motion of the bond.

The parameters k and r_{eq} are characteristic of each pair of atoms. Foundation of MM is the concept of *atom types* and *transferability*: parameters are obtained from experimental data or quantum mechanical calculations for a small subset of molecules and are assumed to be transferable to other molecules with the same atom types. Atom types are classification based on the element and the bonding environment. The set of atom types and their relative parameters form the *force-field*.

Force-fields also contain a term that describes the bending motions of atoms. Even in this case the motion is a fluctuation around an equilibrium position, so it is modelled with a harmonic function:

$$U_{bend_{(\theta)}} = \frac{1}{2}k_\theta\left(\theta - \theta_{eq}\right)^2$$

where ϑ is the angle formed among three atoms, ϑ_{eq} the reference value for the angle and k_ϑ the force constant that characterize the bending motion. The accuracy of both the bond-stretching and the angle-bending terms can be improved by incorporation of higher order terms to the harmonic form of the potential.

The last motions that complete the description of the possible internal degrees of freedom of a molecules are torsions. Torsional motions cannot be accounted for by a harmonic potential because they must be periodic. For this reason, the simplest torsional function implemented in force fields are cosine functions in the form:

$$U_{tors_{(\varphi)}} = A(1 + cos(n\varphi - \varphi_s))$$

This is a periodic function with minima positions determined by the φ_s value, periodicity 360°/n and amplitude A. Torsional potential energy anyway can also assume different forms with maxima at different height (Figure 2.1). To account for this different shaped energy function, more complex expressions of the dihedral energy can be applied as the *Ryckaert-Bellemans potential:*[90]

$$U_{tors_{(\varphi)}} = \sum_{n=0}^{5} C_n(cos(\varphi - \pi))^n$$

Where the C_n are six different parameters. An alternative way for the definition of the torsional potential energy is the *Fourier function*:

$$U_{tors_{(\varphi)}} = \frac{1}{2}[C_1(1 + cos(\varphi)) + C_2(1 - cos(2\varphi)) + C_3(1 + cos(3\varphi)) + C_4(1 - cos(4\varphi))]$$

where C1, C2, C3, C4 are parameters that can be optimized to fit the experimental or quantomechanical data (Figure 2.1).

In addition to the internal degree of freedom, the non-bonding interactions between pairs of atoms are included in the potential. They are treated with two terms describing electrostatic and van der Waals interactions.

Electrostatic interactions between atoms i and atom j are treated with the *Coulomb's law*:

$$U_{coul_{(r_{ij})}} = f\frac{q_i\, q_j}{\varepsilon_r\, r_{ij}}$$

where f is the electric conversion factor ($\frac{1}{4\pi\varepsilon 0}$), q_i and q_j are the net atomic charges of atoms i and j, ε_r is the dielectric constant of the medium and r_{ij} is the distance between atom i and atom j.

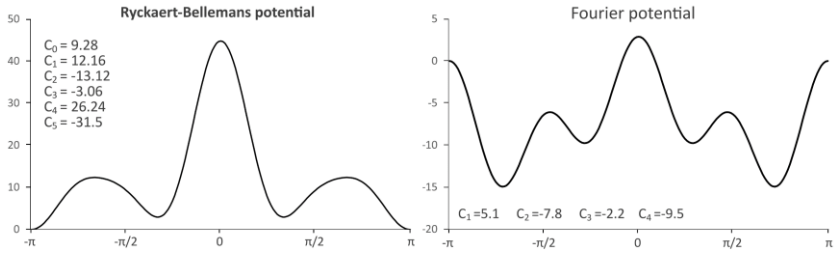

Figure 2.1: Alternative definition of dihedral potential that allow a better fitting to experimental/quantomechanical data. An example of Rickaert-Bellemans potential (left panel) and Fourier potential (right panel).

The shape of the Coulomb potential for atoms of the same (repulsive) and opposite (attractive) charges is illustrated in Figure 2.2. Coulomb potential for two atoms with opposite charge has its minimum at distance = 0 because it does not take into account the radius of the particles. This is introduced with the van der Waals term which is usually modelled with a *Lennard-Jones potential* function:

$$U_{LJ_{(r_{ij})}} = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$

where ε_{ij} is the depth of the potential well and σ_{ij} is the finite distance at which the inter-particle potential is zero (Figure 2.2).

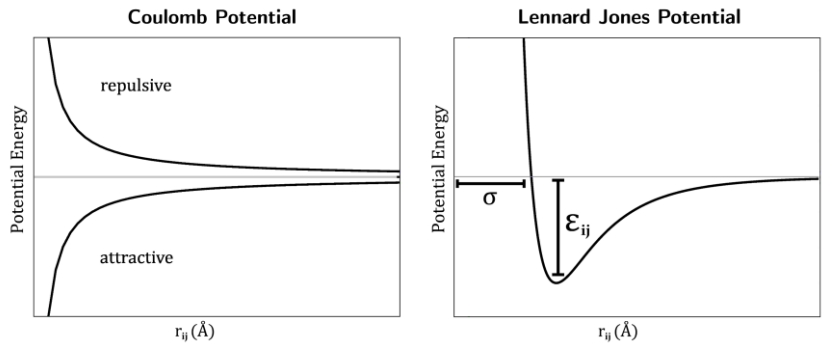

Figure 2.2: Shape of the Coulomb (left panel) and Lennard Jones (right panel) potential functions.

Atomic charges and Lennard-Jones parameters are properties of each atom type (or of each pair of atom types) and are part of the specific force-field.

In the *Amber* force-fields used in the following studies, Lennard-Jones parameters are derived from simulations of liquids for homo-interactions, and the cross terms involving different atom types i and j are evaluated according to the Lorentz/Berthelot mixing rules:[91]

$$\sigma_{ij} = \frac{1}{2}\left(\sigma_{ii} + \sigma_{jj}\right)$$

$$\varepsilon_{ij} = \sqrt{\varepsilon_{ii} \cdot \varepsilon_{jj}}$$

in which arithmetic average is used to calculate σ_{ij} while a geometric average is used to calculate ε_{ij}. Charges are instead derived from the electrostatic potential computed from gas phase quanto-mechanical calculations at the Hartree-Fock level with the 6-31G* basis set (HF/6-31G*), with the Restricted Electrostatic Potential (RESP) method.[92]

2.2 Energy Minimization

Given the expression of energy as a function of the atoms coordinates, it is possible to calculate the energy associated to a particular set of coordinates (conformation). In molecular modelling one is usually interested in low energy conformations, corresponding to stable states of the system, but given the high dimensionality of the function describing the energy as a function of the coordinates of atoms, it is impossible to characterize the potential energy landscape analytically and determine the conformations at lower energy. Energy minimization allows to reach the nearest low energy stationary point on the *Potential Energy Surface* (PES), optimizing the geometry of the system according to the force-field energy.[93] This type of

calculation lead to a local minimum, but does not ensure the achievement of the global minimum.

A stationary point is characterized by the first derivatives of the function with respect to all the variables equal to zero. Energy minimization approaches are classified based on the order of derivatives considered to reach a minimum. Highest the order of the derivatives, more time is required for the computation. Simplest methods are zero order and do not require the derivatives calculation, anyway with modern computer first or second order methods are usually employed.

The *steepest descent* approach, for example, is a first order method that follows the general algorithm:

```
k=0
while ∇f(x_k) ≠ 0:
        calculate the descent direction p_k: -
∇f(x_k)
        calculate the step size α_k
        x_{k+1} = x_k + α_k p_k
        k=k+1
end
```

where $\nabla f(x_k)$ is the gradient of $f(x)$ in the x_k point. In the case of a *stationary* method, α_k is constant, while for *dynamical* methods it is computed at every step, optimizing its value with an analytical or approximate approach. This method is a fast method for a local optimization but can have troubles in reaching convergence in such cases. In particular, it can require many steps to converge when it approaches the minimum.

An alternative first order method is the *conjugate gradient* method which assure the convergence in n step (where n is the number of dimension of the system) but can be used as an iterative method setting a stopping criterion. This kind of method is more computationally demanding, but do not suffer of the oscillating

problems near the minimum present in the steepest descent method. For these reason, a common strategy used to optimize the minimum search is to first perform a steepest descent minimization with a limited number of steps, and then switch to the more demanding conjugate gradient when the algorithm is approaching the minimum.

An example of second order methods is the *Newton Raphson* method in which, given a starting point x_0, the first derivative of the function is expanded with a Taylor series around x_0, truncated to the second order:

$$f'_{(x_0+\delta)} = f'_{(x_0)} + f''_{(x_0)}\delta$$

If δ is the displacement needed to reach the minimum point, then $f'_{(x_0+\delta)}=0$ and we obtain:

$$\delta = -\frac{f'_{(x_0)}}{f''_{(x_0)}}$$

This relation expresses the displacement δ needed to reach the minimum of the function in one step but given that the Taylor series was truncated to the second order it works only for second order functions at most. For function of superior order, the process can be iterated. This method is effective when the starting point is near to the minimum or when the function has a harmonic behaviour near the minimum.

The methods here presented are *downhills* methods which means that they only allow to modify the coordinates in a way that decrease the energy of the system, but they cannot overcome barriers between different basins. These methods are usually employed for geometrical optimization of molecular structures aimed at removing clashes and bad contacts between atoms.

2.3 Molecular Dynamics Simulations

In MD, successive configurations of the system are generated by integrating Newton's law of motion.[94] The result is a *trajectory* that specifies how positions and velocities of the particles in the system vary with time. The three laws of motions in their original form are:[95]

> *Lex I: Corpus omne perseverare in statu suo quiescendi vel movendi uniformiter in directum, nisi quatenus a viribus impressis cogitur statum illum mutare.*
>
> *Lex II: Mutationem motus proportionalem esse vi motrici impressae, et fieri secundum lineam rectam qua vis illa imprimitur.*
>
> *Lex III: Actioni contrariam semper et aequalem esse reactionem: sive corporum duorum actiones in se mutuo semper esse aequales et in partes contrarias dirigi.*

These statements are at the foundation for classical mechanics and describe the relationship between a body and the forces acting upon it, and its motion in response to those forces. The meaning of the laws can be summarized as follow:

> *First law: In an inertial frame of reference, an object either remains at rest or continues to move at a constant velocity, unless acted upon by a force.*
>
> *Second law: In an inertial reference frame, the vector sum of the forces F on an object is equal to the mass m of that object multiplied by the acceleration a of the object: $F = ma$.*
>
> *Third law: When one body exerts a force on a second body, the second body simultaneously exerts a force equal in magnitude and opposite in direction on the first body.*

Given the particles positions, the forces acting on the system at time t can be calculated from force-field functions and the positions of the particles at time $t+dt$ can be obtained by integration of the differential equations associated to the Newton's second law:

$$\frac{\delta^2 x_i}{\delta t^2} = \frac{F_{x_i}}{m_i}$$

This equation describes the motion of a particle i of mass m_i, where F_{x_i} is the force acting on the particle in the x_i direction. Under the influence of a continuous potential, the motion of all particles is coupled together, giving rise to a many-bodies problem that cannot be solved analytically.

Integration of the equations of motion

The equations of motion are thus integrated using a *finite difference method*. The essential idea of these approaches is that integration is broken down into many small stages, each separated in time by a fixed time step δt. If δt is small enough, forces may be considered constant during this time interval allowing the calculation of new position at time $t+\delta t$. There are many algorithms for integrating the equations of motion using finite difference approaches. All algorithms assume that positions, velocities and accelerations can be approximated as Taylor series expansions:

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \delta t \boldsymbol{v}(t) + \frac{1}{2}\delta t^2 \boldsymbol{a}(t) + \frac{1}{6}\delta t^3 \boldsymbol{b}(t) + \frac{1}{24}\delta t^4 \boldsymbol{c}(t) + \cdots$$

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

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

where r is the position, v is the velocity (the first derivative of the positions with respect to time), a is the acceleration (the second derivative), b and c are respectively the third and fourth derivative. Two of the most used methods for integrating the equation of motion are the *Verlet* and the *leap-frog* algorithms.

The Verlet algorithm[96] is based on the relationships:

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \delta t\boldsymbol{v}(t) + \frac{1}{2}\delta t^2\boldsymbol{a}(t) + \cdots$$

$$\boldsymbol{r}(t-\delta t) = \boldsymbol{r}(t) - \delta t\boldsymbol{v}(t) + \frac{1}{2}\delta t^2\boldsymbol{a}(t) - \cdots$$

which express the position of particles at time $t+\delta t$ and $t-\delta t$. Adding these two equations, truncated at second order, gives:

$$\boldsymbol{r}(t+\delta t) = 2\boldsymbol{r}(t) - \boldsymbol{r}(t-\delta t) + \delta t^2\boldsymbol{a}(t)$$

which allows to calculate the position of the particle at time $t+\delta t$ without explicitly computing velocities. These can be estimated at the half step by:

$$\boldsymbol{v}(t+½\delta t) = \frac{\boldsymbol{r}(t+\delta t) - \boldsymbol{r}(t)}{\delta t}$$

Verlet algorithm has some drawbacks as it provides the particle positions adding a small term ($\delta t^2\ \boldsymbol{a}(t)$) to the difference of two much larger terms, $2\boldsymbol{r}(t)$ and $\boldsymbol{r}(t-\delta t)$, which may lead to a loss in precision.

The leap-frog algorithm[97] instead uses the following relationships:

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \delta t\boldsymbol{v}(t+½\delta t)$$

$$\boldsymbol{v}(t+½\delta t) = \boldsymbol{v}(t-½\delta t) + \delta t\boldsymbol{a}(t)$$

which allow calculation of the particle positions according to velocities at time $t+½\delta t$, which in turn are calculated from velocities of the previous step ($t-½\delta t$) and acceleration at time t. This *"leap-frog"* behaviour (Figure 2.3) gives the name to the method. Velocities at time t can also be obtained as mean of velocities at time $t-½\delta t$ and $t+½\delta$.

Both these methods are not *self-starting* methods and require the generation of random velocities from a *Boltzmann* distribution at the first step.

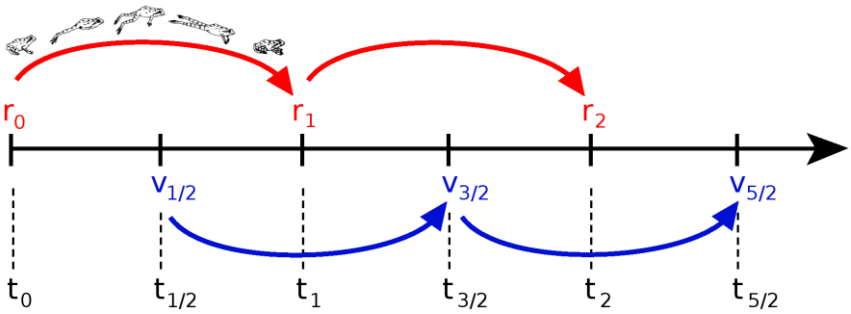

Figure 2.3: Schematic representation of the leap-frog algorithm: velocities 'leap-frog' over positions and positions leap over velocities and so on. (Image modified from ref [94])

As discussed before, the approximation of considering the forces constant within the time step δt is only valid if the time step used is small. But how much the time step should be small? If it is too small, trajectory will progress slowly, thus covering only a limited portion of the phase space, while if it is too large, instabilities may arise in the integration algorithm due to high energy overlaps between atoms. These extreme situations are represented in Figure 2.4. A good rule for determining the time step is that it should be one-tenth the time of the shortest period of motion. In biomolecular simulations, the highest-frequency motion is the stretching of bonds involving hydrogen atoms which vibrate with a period of about 10 fs. Following the above rule implies the use of a 1 fs time step. However, such bond stretching motions are usually of relatively little interest and have

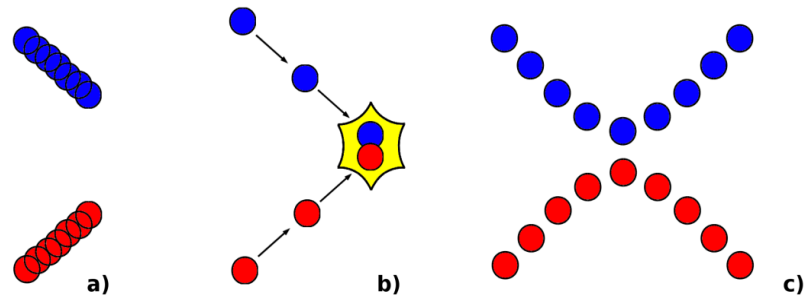

Figure 2.4: With a very small time step (a) phase space is covered very slowly; a too large time step gives instabilities (b); with an appropriate time step simulation proceeds fast and collisions occur smoothly (c). (Image modified from ref [94])

minimal effect on the overall behaviour of the system. To allow the use of a larger time step, usually algorithm like SHAKE[98] or LINCS[99] are used to "freeze out" such vibrations by constraining the appropriate bonds to their equilibrium values.

Boundary effects

Simulate a finite system implies the use of walls limiting the diffusion of molecules in the infinite space. The correct treatment of boundaries and boundary effects is crucial to simulation methods because it enables macroscopic properties to be calculated from simulations using relatively small numbers of particles. This issue is particularly relevant when simulations are performed by explicitly treating the solvent molecules (*explicit solvent*). Walls can produce artefacts, especially in proximity of them, and a method to avoid it is using a *periodic boundary condition* (PBC) approach. In this strategy, a unitary cell of the system is replicated in all directions forming periodic images of the same box. During simulation, all boxes are identical and a particle leaving the central box is thus replaced by the same particle entering from the opposite side of the box. In this way, the number of particles within the cell is constant (Figure 2.5).

The unitary cell must have a shape that can be replicated in the 3D space forming a lattice with no hole. The most common choices are cubic or truncated octahedron boxes. The latter is indicated in case of globular proteins because it approximates a sphere shape and decreases the number of water molecules to be treated, thus reducing the computational time.

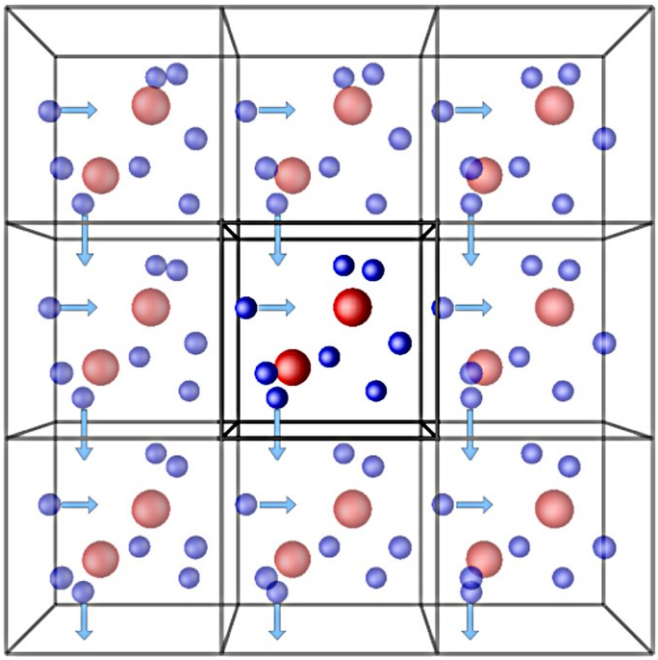

Figure 2.5: Schematic representation of periodic boundary conditions: multiple repetitions in three dimensions of the same box imply that when a particle escapes from one side of the box, a copy enters from the opposite side.

Treatment of non-bonded interactions

Systems studied in biomolecular simulations are often composed of thousands to millions of atoms. Such systems require the computation of a large number of interactions (the total number of non-bonded interactions is proportional to the square of the number of atoms) many of which provide a relatively small contribution to the energy of the system due to the long distance between atoms. The most popular way to deal with the non-bonded interactions is to use a *non-bonded cut-off*, excluding computation of electrostatic and van der Waals interactions for atoms which distance is greater than the selected cut-off. To do that, a neighbour list is calculated and updated every 10-20 steps and the distances are calculated at every step only for the atom neighbours, thus avoiding the computation of the

pairwise distances every step. However, the use of a cut-off introduces a discontinuity in both the potential energy and the force near the cut-off value. There are several approaches that can handle the effects of this discontinuity. One approach is to modify the non-bonded potentials by a function that shifts the force to zero at the cut-off. The purpose of this method is to replace the truncated forces by forces that are continuous and have continuous derivatives at the cut-off distance. For coulomb interactions, the use of a *reaction field* or a proper long-range method such as *Particle Mesh Ewald* (PME) is usually a better choice given that the energy values of these interactions at the cut-off distance are not completely negligible. In the reaction field method, electrostatic energy is calculated explicitly within the given cut-off and, above this, it is approximated using a dielectric continuum.[100] Instead, the PME method uses a summation in the Fourier space for the long-range part which quickly converges in the Fourier reciprocal space.[101,102] All these methods allow to reduce the number of non-bonded interactions to be computed, speeding up the simulations. An additional consequence of using a cut-off is that the *minimum image convention* can be applied: each atom should not see the same molecule twice through the periodic images to avoid artefacts. This imply that the box dimensions should be carefully set to respect the minimum image convention according to the non-bonded cut-off.

Constant temperature and constant pressure dynamics

While direct use of molecular dynamics, by integration of the Newton's equation of motion, gives rise to the NVE (constant number of particles, volume and energy) ensemble (microcanonical ensemble), most quantities that we wish to calculate are actually from constant temperature ensembles. The most common alternative ensembles are

the constant temperature and volume (NVT) ensemble, also referred to as the canonical ensemble, and the constant temperature and pressure (NPT) ensemble.

There are several ways of control the temperature of the system during the simulations such as the *Berendsen temperature coupling* or the *velocity rescale*. The Berendsen algorithm mimics weak coupling with first-order kinetics to an external heat bath with given temperature T_0, scaling the velocity of each particle at every step.[103] Absolute temperature T can be computed using:

$$T = \frac{K}{\frac{1}{2}N_{df}k}$$

Where k is Boltzmann's constant, N_{df} is the number of degrees of freedom and K is the total kinetic energy that can be computed as:

$$K = \frac{1}{2}\sum_{i=1}^{N} m_i v_i^2$$

The effect of the Berendsen algorithm is that a deviation of the system temperature from T_0 is slowly corrected according to:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau}$$

which means that a temperature deviation decays exponentially with a time constant τ. Usually the parameter that is requested to be set is instead τ_T:

$$\tau_T = \frac{\tau N_{df} k}{2C_v}$$

where C_V is the total heat capacity of the system, k is Boltzmann's constant, and N_{df} is the total number of degrees of freedom. τ_T accounts for the lower temperature change due to the redistribution of velocity scaling between kinetic and potential energy.

The velocity-rescaling thermostat[104] is essentially a Berendsen thermostat with an additional stochastic term that ensures a correct kinetic energy distribution by modifying it according to

$$dK = (K_0 - K)\frac{dt}{\tau_T} + 2\sqrt{\frac{KK_0}{N_f}}\frac{dW}{\sqrt{\tau_T}}$$

where K is the kinetic energy, N_f the number of degrees of freedom and dW a Wiener process. There are no additional parameters, except for a random seed. This thermostat produces a correct canonical ensemble and still has the advantage of the Berendsen thermostat: first order decay of temperature deviations and no oscillations.

In the same spirit as the temperature coupling, the system can also be coupled to a pressure bath with different pressure coupling strategies such as the *Berendsen* approach.[103] In the same way as the thermostat, the Berendsen pressure coupling algorithm rescales the coordinates and box vectors, which has the effect of a first-order kinetic relaxation of the pressure towards a given reference pressure P_0:

$$\frac{dP}{dt} = \frac{P_0 - P}{\tau_P}$$

The scaling is usually isotropic, even if in some cases semi-isotropic pressure coupling (which is isotropic in the x/y directions and anisotropic for the z direction) may be used such as simulation of membrane proteins.

Setting up a MD simulation

A good rule to start a MD simulation is that the initial conformation is in a stable low energy situation. This may not be true since usually biological systems such as proteins are obtained from X-ray structures which resolution may be low and may have missing

parts that have to be modelled. Moreover, solvent molecules are placed around solute by a computer program and it should be allowed to equilibrate around the solute. To let the system reaches an equilibrated low energy conformation a pre-production stage is usually performed in which the system is first minimized, and then simulated at increasing temperature applying restraints on critical parts (usually protein backbone).

2.4 Hints of Statistical Mechanics

Statistical mechanics is a branch of physics that explores the connection between the microscopic behaviour and the macroscopic observables.[105,1] The *miracle* accomplished by statistical thermodynamics is indeed that properties relevant to actual laboratory experiments can be described accurately using a ridiculously small number of particles, simulated over times that are far from being macroscopic. The first statement that introduces to the world of the statistical physics developed by Gibbs is that the average value of any observable A, function of positions and momenta of the system (Γ) in the phase space is given by an integral of the form:

$$\langle A \rangle = \int d\Gamma \, w(\Gamma) \, A(\Gamma)$$

where $w(\Gamma)$ are *statistical weights* that characterize the ensemble. But what is the ensemble? An ensemble is an idealization consisting of a large number of virtual copies of a system, each of which represents a possible microstate. When performing experiments, the ensemble is given by all the microstates of the system within the sample analysed, and the macroscopic observable measured is the average value of all the microstate within the sample as reported in the above equation. In MD simulations, anyway, it is not possible simulate a system with the dimension of the sample used in

experiment. What is usually done is producing the trajectory of a single virtual copy of the system, that sample different microstates during the simulation time. In the approximation of a long simulation where all the possible microstates have been extensively sampled, the time and the configurational ensemble correspond (Figure 2.6).

In other words, considering the simulation of a molecule in solution, the equilibrium ensemble exactly represents the fraction of time a molecule spends in different configurations. Configurations appear more commonly in the equilibrium ensemble because they take up more time in the life of any single molecule. This lead to the so called *ergodic hypothesis*: average properties calculated over time or over the statistical ensemble are the same.

Consider now an isolated system A_0 composed by a small microscopic system A and a much larger macroscopic system A' that can be considered as a reservoir. The number of states accessible to the A_0 system is given by:

$$\Omega_0 = \Omega_A \Omega_{A'}$$

where Ω_A is the number of microstates of the microscopic system A and $\Omega_{A'}$ is the number of microstates of the reservoir A'. Studying the system A one could be interested in a specific macrostate characterized by energy E_s, volume V_s and number of particle N_s. The

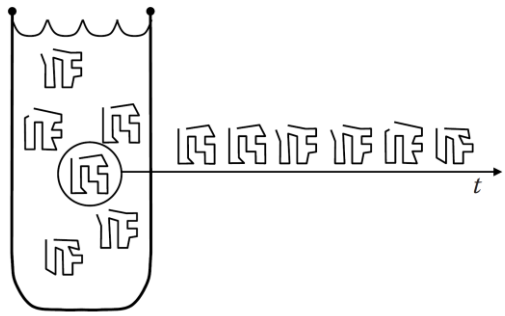

Figure 2.6: Schematic representation of a test tube containing an ensemble of molecules in different configurations, and an ensemble derived from the time evolution of a single molecule. The two ensembles correspond for long observation times. (Image modified from ref [1])

probability of finding the subsystem A in this macrostate is thus proportional to the number of states:

$$\Omega_0(E_0, E_s, V_0, V_s, N_0, N_s) = \Omega_A(E_s, V_s, N_s)\, \Omega_{A'}(E_0 - E_s, V_0 - V_s, N_0 - N_s)$$

where the reservoir *A'* has energy E_0-E_s, volume V_0-V_s and number of particles N_0-N_s.

In the *Boltzman* definition, entropy is a measure of the number of possible microscopic states (or microstates) of a system in thermodynamic equilibrium, and can be expressed as:

$$S = k_b \ln \Omega$$

Imaging the small subsystem A as being *taken out* of *A'*, the change in the entropy of the reservoir can be expressed by the equation:

$$dS' = \left(\frac{dS'}{dE'}\right)_{V',N'} dE' + \left(\frac{dS'}{dV'}\right)_{E',N'} dV' + \left(\frac{dS'}{dN'}\right)_{E',V'} dN'$$

It is possible to demonstrate that:

$$\left(\frac{dS'}{dE'}\right)_{V',N'} = \frac{1}{T'} \qquad \left(\frac{dS'}{dV'}\right)_{E',N'} = \frac{P'}{T'} \qquad \left(\frac{dS'}{dN'}\right)_{E',V'} = -\frac{\mu'}{T'}$$

Where T is temperature, P is pressure and μ is the chemical potential.

The changes in E, V and N are:

$dE' = -E_s \quad dV' = -V_s \qquad dN' = -N_s$

where E_s, V_s and N_s are the energy, volume and number of particles removed from A'.

This means that the change in entropy of the reservoir can be expressed as:

$$dS' = S' - S'_0 = -\frac{1}{T'} E_s - \frac{P' V_s}{T'} + \frac{\mu' N_s}{T'}$$

where S' is the entropy of the reservoir after the system has been taken out, and S_0' is the initial entropy of the system. Expressing the equation in terms of number of microstates accessible to the reservoir:

$$k_B\, ln(\Omega') - k_B\, ln(\Omega_0') = -\left\{\frac{1}{T'}E_s + \frac{P'V_s}{T'} - \frac{\mu' N_s}{T'}\right\}$$

$$ln\left(\frac{\Omega'}{\Omega_0'}\right) = -\left\{\frac{1}{k_B T'}E_s + \frac{P'V_s}{k_B T'} - \frac{\mu' N_s}{k_B T'}\right\}$$

$$\Omega' = \Omega_0'\, e^{-\beta'(E_s + P'V_s - \mu' N_s)}$$

where $\beta = \frac{1}{k_B T'}$ is called *temperature parameter*.

The probability that a certain configuration occurs is related to the number of microstates associated with that particular configuration. Thus the probability of finding the subsystem A with an energy E_s, a volume V_s, and number of particles N_s is given by:

$$P_s = Ce^{-\beta'(E_s + P'V_s - \mu' N_s)}$$

where C is a constant independent of the energy, volume and number of particles in the subsystem. In this expression, the primed quantities explicitly indicate which parameters are associated with the reservoir, but this distinction is often neglected, implying that macroscopic parameters are referring to the parameters of the reservoir in contact with the subsystem, not the subsystem itself. In the following, the primes will be dropped.

In the canonical ensemble the number of particles, the volume and the temperature of the system is fixed (NVT). For this case the probability function can be written as:

$$P_s = Ce^{-\beta E_s}$$

The exponential factor $e^{-\beta E_s}$ is known as the *Boltzmann factor*, and the corresponding probability distribution is called *Boltzmann distribution*. The Boltzmann distribution gives the probability of finding a small subsystem of fixed volume with a particular energy E_s

when the subsystem is in contact with a larger reservoir at constant temperature T. Since the sum of all probabilities must equal to unity, the constant of proportionality C, which is independent of E_s, can be determined using the normalization condition:

$$C = \frac{1}{\sum_s e^{-\beta E_s}} = \frac{1}{Z}$$

where the sum is taken over all energy states accessible to the subsystem. This sum is known as the partition function Z. The partition function is an extremely useful function in statistical mechanics, because a knowledge of the partition function is all that is needed to derive all the pertinent macroscopic parameters of a system. Thus, the probability of finding a subsystem of fixed volume with a particular value of the energy E_s can be expressed in terms of the partition function as:

$$P_s = \frac{e^{-\beta E_s}}{Z}$$

This last equation is one of the most powerful equations in statistical physics. It relates the probability of a microstate with its energy. Barrier crossing events can be analysed in the framework of this definition. Imagining the system starting from state A, (Figure 2.7a) then the relative probability of the transition state (the barrier top) is given by:

$$\frac{e^{-\frac{U^{\ddagger}}{k_B T}}}{e^{-\frac{U_A}{k_B T}}} = e^{-\frac{\Delta U}{k_B T}}$$

Where $U^{\ddagger}$ is the barrier top energy, U_A is the energy of the A minimum and ΔU is the activation energy $U^{\ddagger} - U_A$. The $e^{-\frac{\Delta U}{k_B T}}$ term represent the Arrhenius factor which contains important information about relative probabilities at the equilibrium, but also about the dynamics of the system. The only way a transition can occur is indeed if the barrier top is reached. It is thus clear that the Arrhenius factor

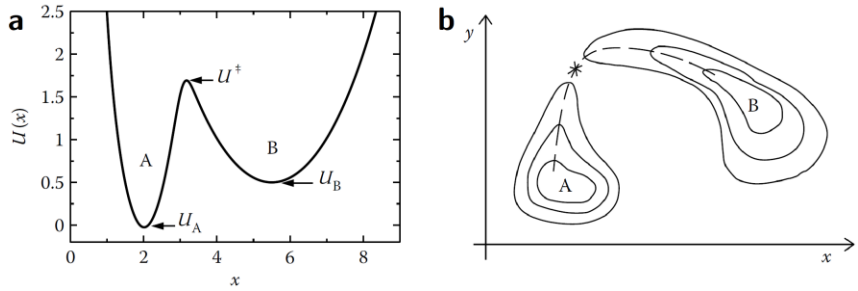

*Figure 2.7: Illustration of barrier crossing: a) a simple one-dimensional model that contains an energy barrier between state A and state B. b) A 2D model in which the transition state is indicated with the * symbol. PMF projections into either x or y clearly yield the wrong transition state. (Image taken from ref [1])*

must enter the dynamical description. The rate constant k is indeed proportional to the Arrhenius factor:

$$k \propto e^{-\frac{\Delta U}{k_B T}}$$

Qualitatively, the Arrhenius factor tells us that:

- the larger the barrier height (ΔU), the slower the transition rate;
- for a given barrier, we expect the rate to increase with increasing temperature.

This information is of fundamental importance for the understanding of conformational sampling in MD simulations.

From the partition function it is also possible to derive all the macroscopic properties of the system:

- the average value of the total energy:

$$\langle E \rangle = \sum_s E_s P_s = \frac{1}{Z} \sum_s E_s \, e^{-\beta E_s} = -\frac{d \ln Z}{d\beta}$$

- the entropy:

$$S = -k_B \sum_s P_s \ln P_s = k_B (\ln Z + \beta \langle E \rangle)$$

- the Helmholtz free energy:

$$F = -k_B T \ln Z$$

The thermodynamic definition of the Helmholtz free energy is: F = E –TS. Thus: $F = \langle E \rangle - T\mathrm{S}$, which is equivalent to:

$$e^{-\frac{F}{k_B T}} = e^{-\frac{\langle E \rangle}{k_B T}} e^{+\frac{S}{k_B}}$$

Since the Boltzmann factor of the free energy gives the probability of a state, $e^{+\frac{S}{k_B}}$ would appear to indicate how many times the Boltzmann factor of the average energy needs to be counted in order to correct for the full probability (that means, for the extent of the state).

Given that the probability of finding a molecular system in one state or the other is determined by the difference in free energy between those two states, it is extremely important being able of calculating the free energy difference between two states. Within the framework of statistical mechanics, a variety of formulae for determining this quantity, or the projection of such a difference along a reaction coordinate or another coordinate in the parameter space, can be derived. The different formulations available are all equivalent within the limit of infinite sampling of the phase space. The most straightforward way to determine the difference in free energy between two states of a system is the so called *direct counting* method[106] in which the number of configurations in the corresponding states are simply counted. The free energy difference between state A and state B can be computed as:

$$\Delta F_{BA} = -k_B T \ln \left(\frac{N_B}{N_A} \right)$$

where N_B and N_A are the number of configurations counted respectively for state A and state B in the simulation. This technique is only appropriate when both states (for example, the bound and unbound configurations) occur with sufficient frequency in the

ensemble to obtain reliable statistics, i.e., when the ΔF_{BA} is small and the barrier that determines the rate of binding and release is also small.

The comprehension of the physics of a process often requires proper representation of the underlying energy. High-dimensional systems anyway are difficult to handle not only because the calculations are hard, but also because of our limited ability to visualize and understand dimensionality higher than three. The trick that allows to visualize the energy of a high dimensional system is using projections. A projection can be performed from an original configuration space of any size down to any smaller size space. As an example, for a two dimensional system it is possible to project the probability in the single x dimension:

$$\rho(x) = \int dy\, \rho(x, y)$$

The statistical mechanical analogue is a free-energy-like quantity called the *potential of mean force* (*PMF*).[4] It is similar to the free energy because the Boltzmann factor of the *PMF* is defined to give the probability distribution on the coordinates of interest. In other words, the *PMF* is the free energy landscape corresponding to the distribution in a subset of coordinates:

$$e^{-\frac{PMF(x)}{k_B T}} \propto \rho(x)$$

Sometimes, *PMF* is referred to as a *free energy profile*. A *PMF* can give a distribution for any variable of the system that depends on the coordinates. For a general variable R (like a distance) which can be determined for any configuration of the system: $R = \hat{R}(\boldsymbol{r^N})$ where $\hat{R}$ is the mathematical function and R is the value it yields, it is possible to calculate the probability distribution, and so the PMF:

$$e^{-\frac{PMF(R)}{k_B T}} \propto \rho(R) \propto \int d\boldsymbol{r^N} \delta\left(R - \hat{R}(\boldsymbol{r^N})\right) e^{-\frac{U(\boldsymbol{r^N})}{k_B T}}$$

In many cases, the *PMF* is incredibly easy to compute, even for complicated variables like distances or RMSDs. A good simulation indeed will sample the desired distribution, which is proportional to a Boltzmann factor. Once a set of configurations has been sampled, it is sufficient to calculate R for every configuration r^N and bin the values to make the histogram. The histogram will therefore be proportional to $e^{-\frac{PMF(R)}{k_B T}}$, and the *PMF* can be obtained from the log of the histogram.

Even if the *PMF* provides one of the few ways to rigorously analyse a high-dimensional system in a way that our low-dimensional minds can understand it, a warning about over-interpretation of *PMF* must be issued. Looking at Figure 2.7b it is tempting to obtain the transition state and barrier height from projections. But in principle it is wrong. The reaction coordinate connecting two states must describe all the essential aspects of a reaction or structural transformation. The free energy along a coordinate may yield a good picture of the transition state and presumably embodies an estimate of the reaction rate via the barrier height but, if important variables that describe the transformation are neglected, the derived physical interpretation may be wrong. In the case of Figure 2.7b if one imagines obtaining a *PMF* projected into either x or y coordinates, the resulting free energy profile clearly yield an incorrect description of the process of interest.

2.5 Coarse-Grained Molecular Dynamics

Computational modelling of molecular mechanisms of biological processes requires models that can reproduce accurately not only the structural and the dynamical properties of all molecular entities involved but also the transient intermolecular interactions in which these entities engage and that modulate their various functional states. This task is often complicated further by the size of the

biological systems involved and by the time scales over which these functional processes occur.

A possible way to extend molecular modelling and bridge it with experimental techniques is to use coarse-graining, *i.e.* to represent a system by a reduced (in comparison with an all-atom description) number of degrees of freedom.[71] Due to this reduction and elimination of fine interaction details, the simulation of a *coarse-grained* (CG) system requires less resources and goes faster than that for the same system in all-atom representation. As a result, an increase of orders of magnitude in the simulated time and length scales can be achieved. A large diversity of coarse-graining approaches is available,[107] ranging from qualitative models to models including chemical specificity. In this chapter the Martini force-field will be presented as it has been used during the PhD work.

The first version of this CG force field, developed for simulation of lipids, was published by the Marrink group in 2004.[108] The name 'Martini' was coined in 2007,[109] while subsequent extension to peptides and proteins[110] was released in 2008, with recent improvements.[111] The Martini model is based on a *four-to-one* mapping, (on average four heavy atoms plus associated hydrogens are represented by a single interaction centre). The four-to-one mapping was chosen as an optimum compromise between computational efficiency and chemical representability.[112] Mapping of water is consistent with this choice, as four real water molecules are mapped to a CG water bead. Ions are represented by a single CG bead, which represents both the ion and its first hydration shell. To represent the geometry of small ring-like fragments or molecules, the general four-to-one mapping approach is too coarse. Ring-like molecules are therefore mapped with a higher resolution of up to two non-hydrogen atoms to one Martini particle.[109] Based on the chemical nature of the underlying structure, a specific particle type with more or less polar character is assigned to each CG bead. The Martini model has four

main types of particle: polar (P), non-polar (N), apolar (C), and charged (Q). Within each type, subtypes are distinguished for their hydrogen-bonding capabilities and their degree of polarity, giving a total of 18 particle types or 'building blocks'.

For proteins, most amino acids are mapped onto single standard particle types (Figure 2.8).[26] The apolar amino acids (*Leu, Pro, Ile, Val, Cys,* and *Met*) are represented as C-type particles, the polar uncharged amino acids (*Thr, Ser, Asn,* and *Gln*) by P-type particles, and the amino acids with small negatively charged side chains as Q-type (*Glu* and *Asp*). The positively charged amino acids *Arg* and *Lys* are modeled by a combination of a Q-type particle and an uncharged particle. The bulkier ring- based side chains are modeled by three (*His, Phe,* and *Tyr*) or four (*Trp*) beads of the special class of ring particles. The *Gly* and *Ala* residues are only represented by the backbone particle. The type of the backbone particle depends on the protein secondary structure; free in solution or in a coil or bend, the backbone has a strong polar character (P type); as part of a helix or strand, the inter-backbone hydrogen bonds reduce the polar character significantly (N type). Proline is less polar due to the lack of hydrogen-donor capabilities. The most appropriate choice of particle

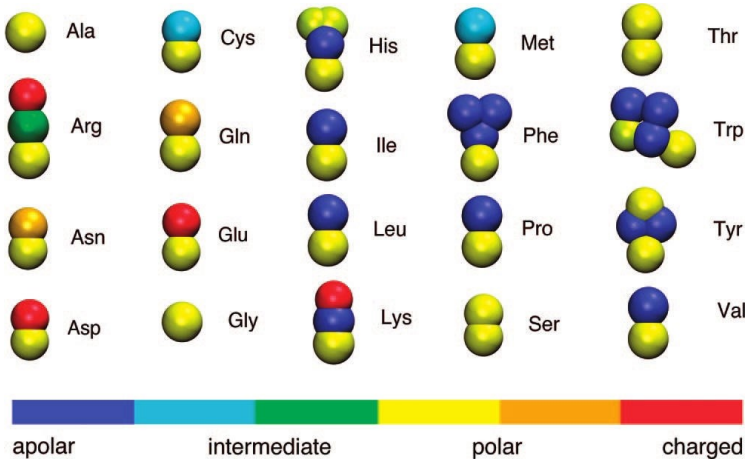

Figure 2.8: Coarse-grained representation of all amino acids. Different colours represent different particle types. (Image taken from ref [110])

types for the amino acids was assessed from a comparison between simulation results and experimental measurements of the water/oil partitioning coefficients of the amino acid side-chain analogues.

In the Martini force field, the non-bonded interactions are treated with Lennard-Jones and Coulomb potentials (see Par. 2.1 Molecular Mechanics). The strength of the interaction is determined by the value of the Lennard-Jones well-depth parameter ε_{ij} that ranges from 5.6 kJ/mol for interactions between strongly polar groups to 2.0 kJ/mol for interactions between polar and apolar groups, mimicking the hydrophobic effect. The effective size of the particles is governed by the Lennard-Jones parameter σ, which has a value of 0.47 nm for normal particle type. For particles in ring-like molecules, $\sigma = 0.43$ nm and the ε_{ij} value is scaled to 75% of the standard value. Lennard-Jones interactions between nearest neighbours are excluded. Charged groups Q bear a charge $\pm$e and interact via a Coulombic energy function, with a dielectric constant of 15 to account for the reduced set of partial charges and resulting dipoles that occur in an atomistic force field. The non-bonded interactions of the Martini model have been parameterized based on a systematic comparison to experimental thermodynamic data. Specifically, the free energy of hydration, the free energy of vaporization, and the partitioning free energies between water and a number of organic phases were calculated for each of the 18 different CG particle types.

Bonded interactions are described by a standard set of potential energy functions common in classical force fields, including harmonic bond and angle potentials, and multimodal dihedral potentials. Proper dihedrals are primarily used to impose secondary structure to the peptide backbone. Improper dihedrals are mainly used to prevent out-of-plane distortions of planar groups. Lennard-Jones interactions between nearest neighbours are excluded. Bonded interactions have been parametrized from structural data (PDB for proteins) or by comparison to atomistic simulations. Importantly, the bonded

parameters depend on the sequence, and are used to stabilize the secondary structure elements of the protein; however, the lack of directional hydrogen bonds prevents realistic folding at the Martini level of coarse-graining. One way to constrain the protein close to a particular state is using elastic network (EN) models. In an EN model, the structure of a macromolecule is described as a network of point masses connected to one another with springs when the distance between the point masses is less than a predefined cutoff distance (R_C). The values of the spring force constant, K_{spring}, and the cutoff R_C characterize the network, that is its rigidity and its extent.

To combine a structure- based coarse-grained model, such as an elastic network, with a physics-based CG molecular force-field to represent a protein, the model named ELNEDIN has been developed.[113] The ELNEDIN approach uses a simple two-parameter (R_c and K_{spring}) elastic network to act as a structural scaffold while the Martini force field directs intermolecular interactions. The model has been optimized with respect to atomistic reference simulations. Figure 2.9 illustrates the various components that make up an ELNEDIN model.

The Martini CG model not only reduce the dimensionality of the system (with the four-to-one mapping the system is described with about 1/10 of the particles needed in an all-atom model), but it also allows to use a longer time step in MD simulations switching from 2 fs of the all-atoms to 20 fs. Moreover, coarse-graining involves modifying the energy landscape to become smoother, which effectively results in more sampling of the energy landscape in a given time period, speeding up the kinetics of the system. This is one of the main advantages of coarse-grained models, but the speed-up is not easily predictable and is not likely to be the same for all degrees of freedom.[114] In Martini, the current best estimate of a semi-universal

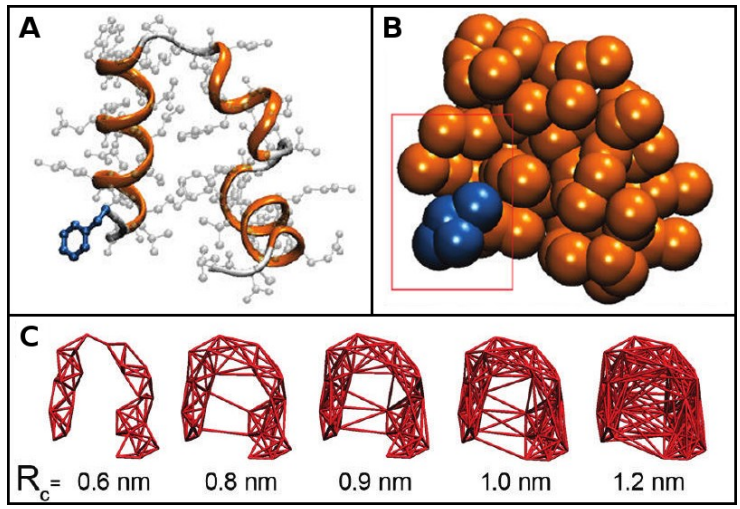

Figure 2.9: Components of the ELNEDIN model of the villin headpiece subdomain. A) Ribbon and ball and stick representations of the protein; B) CPK representation of a coarse- grained model; C) Five elastic network scaffolds at different R_c values (Image taken from ref [113]).

factor of speed-up compared to atomistic simulations is about 4, but this value is strongly dependent on the type of process and molecules studied. Overall the CG representation with Martini force-field allows to sample events happening on the ms timescales, that cannot be investigated with all-atoms simulations.

2.6 Accelerated Molecular Dynamics

Many dynamic events of biological molecules cannot be described by MD simulations using an all-atom description because with this approach often systems remain trapped in potential energy minima separated by high free energy barriers for long simulation times. The dynamic evolution of many molecular systems often occurs through a series of rare events by which the system moves from one potential energy basin to another. It is thus of fundamental importance enhance sampling of rare events maintaining an all-atom description of the system. In the *accelerated molecular dynamics* (aMD) approach the

potential energy landscape is altered by adding a bias potential to the true potential such that the escape rates from potential wells are enhanced; this accelerates sampling and extends the time scale in molecular dynamics simulations. aMD allows to perform unconstrained simulations, i.e. it not involves the introduction of constrains for a specific set of degrees of freedom or collective variables (CVs), that should be determined without a priori knowledge about the topological features of the potential energy landscape.[115] Unlike other methods that involve biasing of the potential, restricting the phase space and knowing the progress coordinate are not requisites of accelerated MD.

The bias potential depends on two parameters: the boost energy limit, E_{lim} , and the tuning parameter α:

$$\Delta U(r) = \frac{\left(E_{lim} - U(r)\right)^2}{\alpha + E_{lim} - U(r)}$$

E_{lim} and α are constants that are set before starting an accelerated MD run. When the potential energy of a system is lower than the boost energy E_{lim}, simulation is performed on the modified potential:

$$U^*(r) = U(r) + \Delta U(r)$$

On the contrary, in regions where the potential energy is equal to or higher than the threshold energy E_{lim}, MD is performed with the original unbiased potential. This form of $\Delta U(r)$ ensures that the derivative $\frac{dU^*(r)}{dr}$ is continuous for any value of $U^*(r)$. How aggressively a system is accelerated depends on the selection of E_{lim} and α. A hypothetical unmodified potential and some potentials obtained by adding various levels of bias are shown in Figure 2.10.

When α tends to infinity the modified potential approaches the original potential, whereas when α tends to zero the potential is flattened toward E_{lim}. The E_{lim} value should be set above the average potential energy of the system obtained from a short conventional

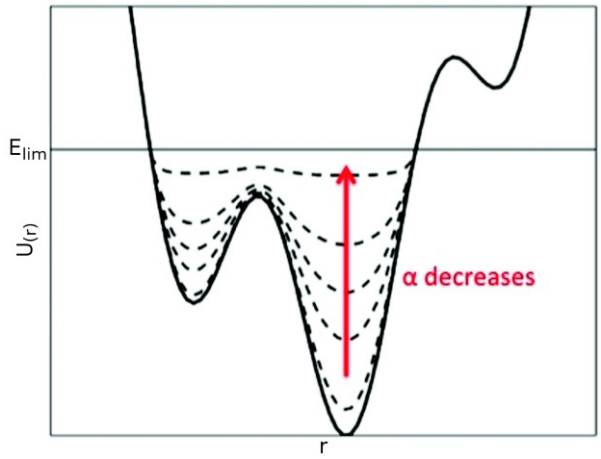

Figure 2.10: Schematic representation of a 1D potential (solid line) and modified potentials $V^*(r)$ (dashed lines) at different α parameter values. (Image taken from ref[116])

unbiased MD after equilibration. Ideally, E_{lim} should be set to a value below the transition state energy. In such way any value of α (except when α is close to zero) will preserve the basic shape of the potential energy surface, maintaining the proportionality between the ratio of the rates of escaping out of a minimum obtained from the unmodified and the modified potentials.

The potential energy terms that compose a force-field have been presented in Par. 2.1 Molecular Mechanics. In principle, the bias potential can be applied to the total potential energy or any of its contributing terms can be boosted separately. In the original version of aMD,[76] the boost potential was applied to the dihedral and the 1-4 nonbonded interaction terms of the force-fields. This approach is referred to as *torsional aMD* and generally it is the method of choice when studying poly-peptides and proteins using implicit solvent models.[117] When proteins are simulated in explicit solvents, the boost is usually applied to the total potential energy[118] (hereinafter referred to as *single boost aMD*), or an extra boost can be applied to the dihedral component[119] (*dual boost aMD*). In case of dual boost aMD,

separate α and E_{lim} parameters have to be set for the total and dihedral boosts.

In principle it is possible to obtain the corrected canonical ensemble averages for the system by simply reweighting each point in the configuration space obtained with the modified potential by the strength of the Boltzmann factor of the bias potential energy, $e^{\beta \Delta U(r)}$, at that particular point.[76] The equilibrium ensemble average value of any observable $A(r)$ taken on the modified potential $U^*(r)$ is given by:

$$\langle A^* \rangle = \frac{\int d\boldsymbol{r}\, A(\boldsymbol{r})\, e^{-\beta U^*(\boldsymbol{r})}}{\int d\boldsymbol{r}\, e^{-\beta U^*(\boldsymbol{r})}}$$

Substituting $U^*(r)$ with $U(r)+\Delta U(r)$, the above expression turns into:

$$\langle A^* \rangle = \frac{\int d\boldsymbol{r}\, A(\boldsymbol{r})\, e^{-\beta U(\boldsymbol{r})-\beta \Delta U(\boldsymbol{r})}}{\int d\boldsymbol{r}\, e^{-\beta U(\boldsymbol{r})-\beta \Delta U(\boldsymbol{r})}}$$

Reweighting the phase space of the modified potential, by multiplying each configuration by the strength of the bias at each position, the corrected ensemble average $\langle A^C \rangle$ is obtained:

$$\langle A^C \rangle = \frac{\int d\boldsymbol{r}\, A(\boldsymbol{r})\, e^{-\beta U(\boldsymbol{r})-\beta \Delta U(\boldsymbol{r})}\, e^{\beta \Delta U(\boldsymbol{r})}}{\int d\boldsymbol{r}\, e^{-\beta U(\boldsymbol{r})-\beta \Delta U(\boldsymbol{r})}\, e^{\beta \Delta U(\boldsymbol{r})}} = \frac{\int d\boldsymbol{r}\, A(\boldsymbol{r})\, e^{-\beta U(\boldsymbol{r})}}{\int d\boldsymbol{r}\, e^{-\beta U(\boldsymbol{r})}}$$

which is equivalent to the equilibrium observable of $A(r)$ on the normal potential. Therefore, the accelerated molecular dynamics simulation method converges to the canonical distribution, and the corrected canonical ensemble average of the system is obtained by simply reweighting each point in the configuration phase space on the modified potential.

Overall, for an aMD simulation, the probability distribution of a selected variable of the system $p^*(A_j)$, in the j^{th} frame, and the boost potential applied at each frame $\Delta U(r)$, it is possible to recover the original canonical ensemble distribution $p(A_j)$ as:

$$p(A_j) = p^*(A_j) = \frac{\langle e^{\beta \Delta U(r)} \rangle_j}{\sum_{j=1}^{M} \langle e^{\beta \Delta U(r)} \rangle_j}$$

where M is the number of frames. The reweighted PMF can be calculated as:

$$F(A_j) = -\frac{1}{\beta} ln\, p(A_j)$$

Recent improvement in the reweighting procedure use a cumulant expansion to the second order to approximate the exponential term.[120] In principle, reweighting using cumulant expansion is able to greatly suppress the energetic noise, since it collectively reweights many data points along the chosen reaction coordinate. The reweight procedure, anyway, only produces accurate results when applied to small systems (proteins with 10-20 residues). For simulations of larger systems, high boost potentials with a broad distribution ($\delta\Delta U \approx 100-200$ kcal/mol) often occur with the current aMD scheme and accurate reweighting remains challenging.[120]

2.7 Metadynamics

Metadynamics method was originally developed by A. Laio and M. Parrinello in 2002,[121] and is based on the addiction of an history-dependent bias to accelerate barrier-crossing events and reconstruct the free energy profile associated to a process of interest. The idea behind this approach is that the bias potential should discourage the sampling of already visited regions of the free energy surface (FES). The algorithm that guides the bias potential can be depicted in a simplified version as follow:[77]

> *Imagine a walker who, during the night, falls into an empty swimming pool. The walls of the swimming pool are too steep for the walker to climb and the complete darkness makes it difficult for him to localize the shallowest point (lowest saddle). In these*

> *conditions the walker will move more easily downhill, and it is rather unlikely that he will find by chance the lowest saddle. His walk in these conditions resembles that performed by microscopic systems in normal molecular dynamics or Monte Carlo: a random walk with a bias in the direction of lower free energy, with a very small probability to explore transition regions (climb out of the pool). In metadynamics, the walker has access to a large source of sand that he can deposit in his current position. The sand will slowly fill the pool.*

The method was indeed depicted as *"filling the free energy wells with computational sand"* (Figure 2.11).

The computational bags of sand mentioned above are instead Gaussian functions that increment an history-dependent bias potential. The algorithm is based on a dimensional reduction, and thus requires the preliminary identification of a set of CVs, *S(x)*, explicit functions of the system coordinate x, which are assumed to be able to describe the process of interest.[77]

In the case of a single CV, the external *metadynamics* potential acting on the system at time t is given by

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

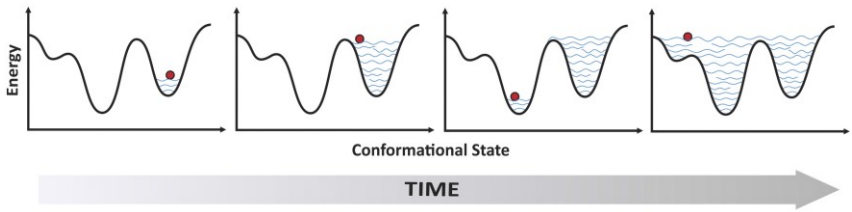

Figure 2.11: Illustration of the metadynamics method described as "filling the free energy wells with computational sand". (Image modified from ref [69])

Where *s(t)=S(x(t))* is the value taken by the CV at time *t*. Three parameters enter the definition of U_G:

- The Gaussian height *w;*
- The Gaussian width *δs;*
- the frequency τ_G at which the Gaussians are added.

These parameters influence the accuracy and efficiency of the free energy reconstruction. Qualitatively, they define the volume of the *'sand'* the walker is depositing. If the Gaussians are large, the free energy surface will be explored at a fast pace, but the reconstructed profile will be affected by large errors. Instead, if the Gaussians are small or are placed infrequently the reconstruction will be accurate, but it will take a longer time.

Width and height are usually defined based on previous simulation runs or modified after with a trial and error approach. A short unbiased MD run is usually performed starting from a hypothesized minimum of the system to determine the shape of the minimum. The Gaussian widths are usually set to one half or one third of the standard deviation of the CVs, computed on this simulation.[122] If one can estimate the barrier height for the process of interest (for example with a previous steered or umbrella sampling MD), the Gaussian height is usually set to about 1% of the barrier height. Finally, the deposition rate should be set in a way that the system has the possibility of equilibrating after the Gaussian deposition, otherwise the following Gaussian will be placed on top of the previous Gaussian.

If the CV is a d-dimensional vector, namely two or more CVs are used at the same time, the metadynamics potential is given by:

$$U_G(S(x),t) = w \sum_{\substack{t'=\tau_G,2\tau_G,\dots \\ t'<t}} e^{-\sum_{\alpha=1}^{d} \frac{\left(S_\alpha(x)-s_\alpha(t')\right)^2}{2\delta {s_\alpha}^2}}$$

and it is necessary to choose a width δsα for each CV. The time required to escape from a local minimum in the free energy surface is determined by the number of Gaussians that are needed to fill the well. This number is proportional to $\left(\frac{1}{\delta s}\right)^d$, where d is the number of CVs used in the system. Hence, the efficiency of the method scales exponentially with the number of dimensions involved. If d is large, the only way to obtain a reasonable efficiency is to use Gaussians with a size comparable to that of the well. From these simple considerations it is clear that the metadynamics works properly only if d is small, and that the quality of the reconstructed free energy is strongly influenced by the parameters w and δs and by the choice of the CVs. Ideally the CVs should satisfy three properties:[77]

- They should clearly distinguish between the initial state, the final state and the intermediates;
- They should describe all the slow events that are relevant to the process of interest.
- Their number should not be too large, otherwise it will take a very long time to fill the free energy surface.

While the first property is quite intuitive, and we already discussed the third property, it may be not clear why all slow motions of the process of interest should be described within the CV. Consider the Z-shaped two-dimensional free energy depicted in Figure 2.12.

If a metadynamics simulation is performed biasing only CV1 and neglecting CV2 the simulation, that is started in basin B, is not able to perform a transition toward A in the due time, and metadynamics goes on overfilling this minimum. A transition is finally observed only when the height of the accumulated Gaussians will largely exceed the true barrier height. This hysteretic behaviour will continue indefinitely without ever reaching a situation in which the free energy grows evenly.

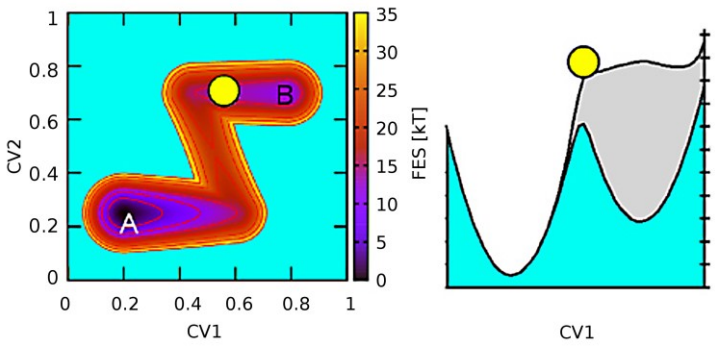

Figure 2.12: The effect of neglecting a relevant CV. Left side: 2D Z-shaped potential. Right side: Using only CV1 causes strong hysteresis effect in the reconstructed free energy. (Image taken from ref [77])

Standard metadynamics presented so far has two well-known problems:[122]

- Its estimate for the free-energy landscape does not converge but fluctuates around an estimate that, at least for simplified systems, can be demonstrated to be unbiased;
- Because it is a flat histogram method, it tries to sample the whole CV space. This can push the simulated system toward states with non-physically high free energy and might drift the simulation toward thermodynamically nonrelevant configurations.

These problems have been recognized and tackled respectively by taking time averages[123] and by using restraining potentials.[124] An alternative method that addresses both the problems in an elegant fashion is *well tempered* (WT) metadynamics.[125] In WT metadynamics, the rule for the metadynamics potential is slightly modified scaling down the height of deposited Gaussians by a factor:

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

where the bias potential has been evaluated at the same point where the Gaussian is centred and ΔT is an input parameter measured in temperature units. This imply that after the initial

filling, Gaussians of different height are added in different regions of the CV space. On top of deep wells, where a sizable bias has been already accumulated, the additional Gaussians have small height, while at the border of the explored region, where the bias is still small, the additional Gaussians have large height. In the long time limit the simulated system should systematically spend more time on the regions where smaller Gaussians are used, that is, on top of deeper wells. This disrupts the flat histogram properties of the method and in turn implies that the sum of the metadynamics potential and of the free energy ($U_G(s)+F(s)$) is no longer encouraged to become flat. As a consequence of the scaled biased deposited, in WT metadynamics the bias does not tend to become the negative of the free energy but is instead a fraction of it. Thus, it only partially compensates existing free-energy barriers by an a priori known scaling factor:

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

This factor is known as *biasfactor* and is an input parameter that modulate the scaling of hills height. It is possible to demonstrate that in the long time limit, the system will explore the biased canonical distribution:[125]

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

Because of the bias potential, the CVs are exploring the canonical ensemble at an effective temperature $T+\Delta T$. It should be underlined that the other microscopic variables are still sampled at the temperature T. WT metadynamics is thus of great advantage as it allows limiting the exploration of the CV space only to regions of reasonable free energy. Indeed, by fixing ΔT according to the height of the typical free-energy barrier for the problem under consideration, one will avoid overcoming barriers that are much higher than that.

Metadynamics is a powerful method particularly suitable to accelerate the study of processes happening on long timescales. One of the main limitations of this method anyway is reaching the convergence, especially if more than one CV is used. To overcome this limitation various approach that speed-up the sampling performing multiple replicas have been developed such as *multiple walkers*,[126] *replica exchange* metadynamics,[127] or *bias exchange* metadynamics.[128] Among these methods the multiple walkers (MW) metadynamics will be discussed as it was used during the PhD work. In MW metadynamics, N_w metadynamics simulations (also referred to as *walkers*) are run concurrently, possibly on different machines. All these simulations contribute to the growth of a unique bias potential, which thus grows at a speed that is N_w times larger than for a single simulation. It has been shown that the resulting error is the same as that expected from a single walker using the same Gaussian height and deposition time.[126,129] This means that when using N_w walkers a filling time acceleration by a factor N_w can be obtained without increasing the error. In MW metadynamics there is no efficiency gain in the computing time, as the same accuracy could be obtained by performing just a single simulation N_w-times longer. However, because the walkers are weakly coupled, this algorithm can be run easily on a parallel machine or even on a weakly interconnected cluster. Only a shared file system is required to allow inter-walker communication. Overall the MW approach is an algorithm that can perform parallel metadynamics simulations to obtain a very fast filling albeit using a slow deposition rate, resulting in a faster convergence.

Another drawback of metadynamics, which is shared with all the methods based on biasing CVs, is that those CVs should be chosen before performing the simulation, and their choice typically affects the accuracy of the final result. However, it is sometimes very useful to compute free energies as functions of CVs that differs from the biased CV. This can be done by an a posteriori reweighting

analysis[130,131] that can recover the *PMF* for a CV not biased during the metadynamics simulation.

Metadynamics is thus an established method aimed at accelerating molecular dynamics simulations and subsequently recovering free-energy landscapes. Its power and flexibility arise from the fact that it allows to exploit the chemical and physical insight of the process under investigation.

"Equations are more important to me, because politics is for the present, but an equation is something for eternity."

Albert Einstein,
as quoted in Brief History of Time: From the Big Bang to Black Holes

CHAPTER 3

MODELLING BINDING WITH LARGE CONFORMATIONAL CHANGES: KEY POINTS IN ENSEMBLE-DOCKING APPROACHES

3.1 Introduction

Ligand-binding is one of the crucial process by which proteins regulate biological mechanisms and protein dynamics play a critical role in it. The induced fit (IF) and conformational selection (CS) models have been proposed to describe the role of conformational dynamics in binding (as discussed in Par. 1.1 Protein dynamics in biomolecular interactions and functions). This topic has received

specific attention for its implication in structure-based drug discovery studies, and a number of computational methods borrowing from the above models have been developed to account for protein flexibility in ligand-binding.[25,26,51,132–134]

A class of methods that relies on the CS model is ensemble docking, that considers ligand docking to predetermined multiple receptor conformations obtained from either experimental data or computational sampling (see Par. 1.1 Protein dynamics in biomolecular interactions and functions).[30,51,135–139] Ensemble-docking approaches are particularly suitable for binding processes that involve large conformational changes, provided that all the relevant protein conformations are considered. MD simulations have been widely used to describe intrinsic protein dynamics, and in recent years, significant increases in computational power have broadened their applicability.[140–142] To improve the efficiency of sampling and to accelerate the crossing of high energy barriers, various enhanced sampling techniques have been proposed (see Chapter 2).[48,143] In particular, accelerated MD[76,144] (see Par. 2.6 Accelerated Molecular Dynamics) has the advantage that it does not require the a priori selection of a reaction coordinate or collective variables; therefore, it can be used when information on the holo structure is missing. In line with this characteristic, this method has been proposed as a valuable tool for ensemble-based docking and screening methods.[51,70] Furthermore, the choice of a reduced set of representative conformations within the trajectory is pivotal to the efficiency of ensemble docking. The set should capture the entire structural diversity of the target with a limited number of significant conformers associated with the different functional sub-states. Cluster analysis has proven to be effective for this aim,[145] and some optimized clustering techniques for the analysis of conformational ensembles have been proposed.[146,147]

In this work, we analyse ligand-binding processes associated with large conformational changes of a protein to elucidate the critical choices in ensemble-docking strategies for effective prediction of the structure of ligand-protein complexes.

We have selected two study cases, in which binding involves different conformational changes of the receptor and different ligand-protein interactions: the acetylcholine binding protein (AChBP)[78] and the allose binding protein (Allose BP).[79,80] The availability of X-ray structures of both the apo and holo forms[148–152] makes the above systems ideal test cases for our study.

The AChBP is a soluble homopentamer homologous to the extracellular N-terminal ligand-binding domain of the nicotinic acetylcholine receptors (nAChRs), transmembrane proteins involved in ion gating at the basis of neuronal response and muscle activity. Detailed structural analysis of the binding of nicotine and other ligands affecting ion flow and neuronal stimulation to the nAChRs is impaired by the large size and the transmembrane spans of these receptors. The AChBP, which shows high similarity in structure and ligand-recognition properties, has been broadly used as a surrogate structure,[78] and several X-ray structures of AChBP-ligand complexes have been resolved. These structures indicate that the pentamer subunits are arranged in a cylinder, with each subunit characterized by an N-terminal helical region and a 10-strand β-sandwich core. Ligands bind the AChBP at the interfaces between each pair of subunits, in a pocket lined with aromatic residues, behind a loop extending from one of the loops known as the “C-loop”. This loop acts as a flexible gate capping the binding site, and its large opening motion governs ligand specificity. This binding process is a prototypical example of non-specific binding of multiple diverse ligands at the same site, modulated by the dynamics of a loop capping the binding pocket.

The Allose BP, similarly to other periplasmic binding proteins, participates in a bacterial ATP-binding cassette transporter system, including a transmembrane permease and an ATP-binding component, which uses the energy of ATP to carry ligands across cytoplasmic membranes. Large conformational changes associated with ligand binding favor productive interactions of the periplasmic component with the permease.[80,153] In particular, Allose BP belongs to the periplasmic sugar-binding proteins subfamily, for which several crystal structures in both the open free state and the closed sugar-bound form have been determined. This system represents an example of large hinge-bending motion of two domains that leads to highly specific ligand-binding.

Our results not only confirm ensemble docking as an effective tool to consider protein flexibility in ligand binding but also identify the most appropriate methodological choices for the two binding mechanisms. By comparing the performances of conventional MD (cMD) and accelerated MD (aMD) simulations, we assess the choice of the best sampling technique to overcome the energy barriers between the minima of the apo system and efficiently sample the rough conformational landscape within the holo minimum region. Then, the most appropriate strategy to select a reduced set of protein conformations relevant to binding within the MD trajectories is evaluated by comparing results obtained using different clustering techniques. Finally, the role of the docking method and the importance of introducing a post-docking refinement stage are verified.

3.2 Methods

Molecular dynamics simulations

Crystal structures for AChBP (PDB ID: 2BYN) and Allose BP (PDB ID: 1GUD) were obtained from the PDB[154] and were pre-processed for simulation using Schrodinger's Protein Preparation Wizard tool:[155] hydrogen atoms were added (replacing the existing ones), all water molecules were removed, disulfide bonds were assigned and residue protonation states were determined by PROPKA[156] at pH = 7.0.

Each system was then prepared for simulation using the tleap module of the AMBER14 package[157,158] and the ff14SB[159] force field, with TIP3P[160] water placed up to 12 Å from the solute and neutralizing the system with Na+/Cl- ions. Where necessary, parametrization of the ligands was performed using the antechamber[161] module of AMBER14, using the Generalized Amber Force Field[162] (GAFF) to assign the atom types and the AM1-BCC method[163,164] to assign charges.

A prior multistage equilibration approach was used to remove unfavourable contacts and provide a reliable starting point for the simulations. The systems were subjected to 1000 steps of steepest descent energy minimization, followed by 1000 steps of conjugate gradient with backbone restraint (20 kcal mol-1 $Å^{-1}$). Subsequently, a 250 ps MD simulation was used to heat the system from 100 to 300 K in the NVT ensemble with backbone restraint lowered to 10 kcal mol^{-1} $Å^{-1}$. Finally, the systems were equilibrated with a 500 ps NPT simulation with low backbone restraint (2 kcal mol^{-1} $Å^{-1}$). All the restraints were removed for the production runs. In all the stages, the temperature was controlled by the Langevin temperature equilibration scheme[165] with a collision frequency of 2.0 ps^{-1} and

pressure targeted to 1 bar using a Berendsen barostat.[103] A time step of 2.0 fs was used, together with the SHAKE[98] algorithm to constrain the bonds connecting the hydrogen atoms. The particle mesh Ewald[101] method was used to treat the long-range electrostatic interactions with the cut-off distances set to 10 Å.

Both cMD and aMD simulations were performed. The aMD[76,144,166] method was used to enhance the sampling of the conformational space. In this method, a continuous non-negative bias potential (ΔV(r)) is added to the system for each point with energy $<$ Elim. The bias potential is defined as:

$$\Delta V(r) = \begin{cases} 0 & , \quad V(r) \geq E_{lim} \\ \dfrac{\left(E_{lim} - V(r)\right)^2}{\alpha + \left(E_{lim} - V(r)\right)}, & V(r) < E_{lim} \end{cases}$$

Therefore, two parameters, α and E_{lim}, must be set. This is referred to as the single-boost approach because all the degrees of freedom are equally subjected to acceleration. An alternative version of the method includes an additional bias potential for the dihedral angles and is referred to as the dual-boost approach. For a complete description of aMD theory, refer to.[76,115,144]

Standard acceleration parameters are determined based on the average potential and the dihedral energies of the equilibrated MD simulations and are calculated according to:[120]

$$E_{d_{lim}} = V_{d_{avg}} + 4\,N_{res} \qquad \alpha_d = \frac{4}{5}\,N_{res}$$

$$E_{p_{lim}} = V_{p_{avg}} + 0.16\,N_{atoms} \qquad \alpha_d = 0.16\,N_{atoms}$$

Where the $V_{p_{avg}}$ and $V_{d_{avg}}$ were derived from a previous unbiased short simulation of 10 ns. In the present, work both the single- and dual-boost versions of the method were used, and standard acceleration parameters and additional alternative sets of parameters were proposed to decrease or increase the magnitude of the boost.

Probability distribution maps

Probability distribution maps were obtained dividing the conformational space of two selected variables in bins and counting the occurrence of frames within each bin. This procedure is similar to the *direct counting* approach discussed in Par. 2.4 Hints of Statistical Mechanics but, in the case of biased simulations, the resulting distribution does not represent the real free energy of the system. In principle a reweighting procedure of the two variables could restore the original free energy topology, but this procedure is challenging for large systems[120,167] and is beyond the aim of this work.

Cluster analysis of the simulated trajectories

Two clustering strategies were applied to select a reduced number of representative conformations from the MD trajectory.

In the first approach, water molecules were removed from each MD frame, and protein structures were aligned to the first frame based on the heavy atoms of the residues in close contact with the known co-crystallized ligand (any residues with at least one atom within 5 Å from the ligand). The same atoms were used within the GROMOS RMSD-based clustering tool[168] to calculate the pairwise RMSD matrix. In the GROMOS algorithm, the neighbours of each data point are defined according to a cut-off distance; the point with largest neighbourhood defines the first cluster medoid. This point and its neighbours are removed, and the algorithm is iterated until all data have been assigned to a cluster. In the present work, the cut-off values were set to obtain a reasonable number (approximately 50) of clusters, but only the centrotypes of the 10 most populated clusters were considered.

The second clustering method, referred to as volumetric clustering in the other sections, was developed by the authors[81]. It is based on

the concept that the ensemble of representative alternative binding conformations should show the maximum difference in the space accessible to the ligands within the binding site. The algorithm first performs a rapid calculation of the unoccupied space within the binding site in each trajectory frame, and then a cluster analysis of the frames is performed according to the shape of the ligand-accessible space. The calculation of the accessible space is based on a grid of points, similarly to the POVME tool.[169,170] A grid encompassing the binding site of the specific system is defined by the user, and the presence/absence of a protein atom within a given cut-off distance from each grid point is verified. A binary vector indicating the occupation state of each point is obtained for every frame. The pairwise distance matrix between all frames is then computed using the Jaccard distance, and the average linkage method is used to obtain a predetermined number of clusters. The centroids of the clusters are used as the set of representative conformations in the MD trajectory. The volumetric clustering approach is conceptually similar to a method recently proposed by Swift and co-workers.[171]

Molecular Docking

Molecular docking was conducted using components of the Schrodinger Suite 2014.[172] All structures were processed prior to docking calculations with the Protein Preparation Wizard tool, as described in the MD simulation subsection. All ligands were prepared using the LigPrep routine, including 3D conformer generation and protonation. Docking was performed with the Glide method that uses a series of hierarchical filters to search for possible locations of the ligand in the binding-site region of a receptor and includes a flexible treatment of the ligands.[173–175] The shape and properties of the receptor are represented on a grid by different sets of fields that provide progressively more accurate scoring of the ligand pose. Glide

SP[173,175] (standard precision) performs exhaustive sampling through initial greedy positioning of the ligand and subsequent optimization, while Glide XP[174] (extra precision) employs an anchor-and-grow sampling approach and also accounts for explicit waters. Both Glide SP and XP use empirical scoring functions designed to maximize separation of compounds with high binding affinity from those with low or no binding ability. The scoring functions include empirically-based functions that account for different interactions (e.g. lipophilic-lipophilic, hydrogen-bond, metal-ligand terms) and also incorporate force-field-based functions that describe Coulomb and van der Waals contributions to the interaction energies.

In this work, docking grids were generated using the default settings, and the available X-ray structures of the ligands were used to define the center of the grid box. Both Glide SP and XP were used for the molecular docking calculations. A rescoring strategy that consists of rescoring the SP poses with the XP scoring function was also tested.

A few docking calculations were performed by scaling the van der Waals radii of the non-polar atoms of the protein by a factor of 0.8. This soft-docking strategy is an initial rough attempt to introduce protein flexibility in docking.[25]

To evaluate the ligand-docking poses, the dRMSD[176] between the ligand-site distances in the docked complex and the corresponding ligand-site distances in the X-ray structure was calculated:

$$dRMSD = \sqrt{\frac{\sum_i \sum_j \left(d_{ij}^x - d_{ij}^m\right)^2}{N}}$$

where x and m are the experimental and docked complexes, respectively; d are the vectors of the distances between the ligand and the binding-site heavy atoms; i and j are the indices of the atoms; and N is the number of comparisons performed. Using this index, the

distance calculation takes into account only the deviation of the relative position of the ligand with respect to the residues belonging to the binding site; it is a better index to evaluate the accuracy of the binding geometry than the RMSD calculated on the absolute positions of the ligand atoms, which neglects the difference in the positions of protein residues in the docked and reference poses.[176]

Tools for trajectory post-processing and analysis

MD trajectories were visually inspected using VMD[177] and analysed using the bio3d R package.[178,179] Images were generated with Pymol.[180]

3.3 Results

Multiple-ligand binding to the acetylcholine binding protein

We selected six X-ray structures of the Aplysia Californica AChBP, showing different arrangements of the C-loop (Figure 3.1): closed conformations in the complexes with alkaloid nicotinic agonists (epibatidine, hepes, lobeline),[148,149] intermediate conformations in the apo form,[148] and more open conformations when bound to the non-competitive nicotinic ligand cocaine[150] and the alkaloid methyllycaconitine antagonist.[148]

We initially assessed the possibility to sample the apo protein energy landscape efficiently to reach all five holo protein conformations by performing extensive MD simulations of the full AChBP pentamer with different sampling protocols, starting from the apo crystal structure (PDB ID 2BYN).

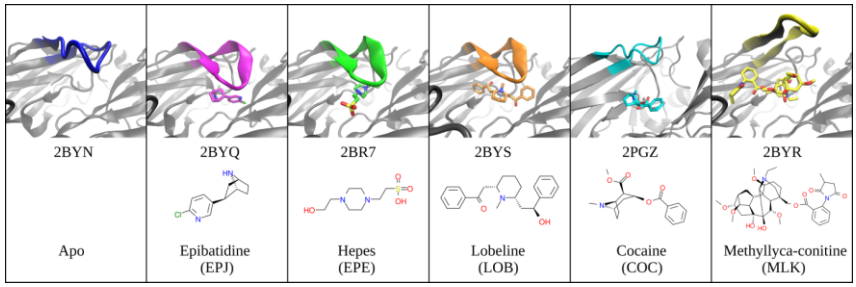

Figure 3.1: Different C-loop conformations in the AChBP X-ray structures here considered. The PDB ID, ligand structure and name (ID) are reported for each structure.

Since the residue positions in the subunits within the pentamer are similar, each pair of subunits forming the binding interface was treated as a replica. At first, 200 ns of cMD simulation was produced for each replica. Additionally, to evaluate whether sampling was improved by the accelerated MD technique (see Par. 3.2 Methods), a 60 ns simulation with the single-boost aMD (all degrees of freedom equally subjected to acceleration) and a 40 ns simulation with the dual-boost aMD (additional bias potential imposed on the dihedral angles) were performed for each replica. The acceleration parameters are reported in Table App. A1.

Plots of the RMSD from each holo structure, computed on the binding-site heavy atoms (binding-site RMSD) and monitored during the simulations, are reported in Figure App. A1, and the percentages of trajectory frames under 1, 1.5 and 2 Å binding-site RMSD were analyzed (Table 3.1). Conformations closer than 2 Å (and, in three cases, closer than 1.5 Å) to each reference structure were broadly sampled in the cMD simulation (Table 3.1a).

Table 3.1: Lowest binding-site RMSD from each AChBP holo structure and percentage of frames under 1, 1.5 and 2 Å RMSD obtained in the different simulations.

	2BYQ	2BR7	2BYS	2PGZ	2BYR
RMSD (Å)			a. cMD		
Lowest	0.80	0.89	1.22	1.02	1.15
% < 1	1.39	0.10	0.00	0.00	0.00
% < 1.5	36.04	31.89	5.51	11.84	4.89
% < 2	54.50	48.42	51.99	56.29	46.72
RMSD (Å)			b. Single-Boost aMD		
Lowest	0.84	1.11	1.24	1.05	1.21
% < 1	0.43	0.00	0.00	0.00	0.00
% < 1.5	12.57	5.84	1.79	9.95	6.47
% < 2	50.56	44.96	46.25	46.05	44.06
RMSD (Å)			c. Dual-Boost aMD		
Lowest	0.92	1.09	1.25	1.29	1.50
% < 1	0.18	0.00	0.00	0.00	0.00
% < 1.5	15.70	7.88	1.50	0.55	0.00
% < 2	55.40	40.68	52.55	10.65	5.78

In all cases, the frame with the lowest RMSD was sufficiently close to the binding-site conformation adopted in the bound protein (approximately 1 Å RMSD) to make it a suitable candidate for docking studies. Compared to cMD, aMD sampled a low percentage of frames close to the holo structures (Table 3.1b, c). With both single- and dual-boost aMD, the number of frames within 1.5 Å from each reference holo structure was less than half the number of cMD frames. Furthermore, fewer frames near the 2PGZ and 2BYR structures were recorded in the dual-boost aMD, indicating difficulty in reproducing the correct open conformations.

To analyze the effectiveness of the different simulations of the apo AChBP in sampling the space related to the fluctuations of the C-loop, the conformational state probabilities were represented in the

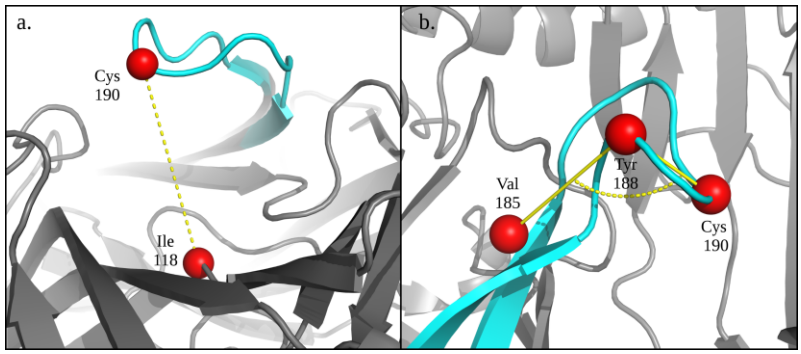

Figure 3.2: Variables selected to describe the sub-space related to the loop motion of AChBP during the simulations: a) distance describing the opening motion of the C-loop; b) angle defining the bending at the tip of C-loop.

subspace defined by two geometric variables (Figure 3.2): the distance between two Cα, to describe the opening motion of the loop; and the angle among three Cα at the tip of the C-loop, to describe the loop bending. The probability distribution map derived from the cMD simulation (Figure 3.3a) shows two frequently sampled zones, corresponding to closed and open arrangements of the loop, with a slightly higher probability for the closed state. Projection of the AChBP X-ray holo structures into this subspace showed that these structures are located in the two frequently sampled areas. In particular, the epibatidine (EPJ), hepes (EPE) and lobeline (LOB) ligands bind a C-loop closed conformation, whereas cocaine (COC)

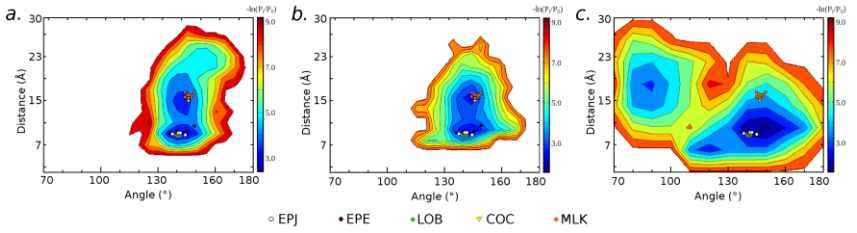

Figure 3.3: Conformational state probabilities from the a) cMD, b) Single-Boost aMD, c) Dual-Boost aMD simulations of the apo AChBP represented in the subspace defined by the two selected variables (Figure S2). The X-ray structures of all the pentamer subunits bound to the five ligands are projected onto the maps. Closed/open states are found at approximately: 9 Å-142°/14 Å-145° by cMD; 10 Å-143°/15 Å-145° by single-boost aMD; and 10 Å-148°/17 Å-90° by dual-boost aMD.

and methyllycaconitine (MLK) choose more open conformations. According to the conformational selection hypothesis, each ligand selects the most appropriate state among multiple pre-existing states in the apo protein free-energy landscape that were effectively sampled in the cMD simulation.

The same map obtained using single-boost aMD (Figure 3.3b) shows that the conformational subspace explored by this simulation (in 60 ns) is similar to that of cMD (200 ns). The dual-boost aMD samples a different region (Figure 3.3c), corresponding to open states in which the loop is bent (low angle values). This explains the low percentage of frames near the open experimental structures (Table 3.1c). The dual-boost aMD rapidly escaped the closed conformation, evolving to an open state distinct from the experimental one.

In addition to the efficiency in sampling the apo protein conformational landscape, the ability to select a reduced number of representative conformations from the MD trajectory is crucial for ensemble-docking applications. Different variables can be considered to cluster distinct macrostates, but for the purpose of ensemble docking, a good clustering protocol should be able to recognize different arrangements of the binding site. Therefore, we employed two different clustering strategies focused on the binding site: the GROMOS RMSD-based algorithm applied to the positions of the binding-site heavy atoms, and the volumetric method based on clustering of the binding-site accessible volumes (see Par. 3.2 Methods).

The binding-site RMSDs from the holo structures for the cluster representatives, obtained by the two approaches, are reported in Table 3.2.

Table 3.2: Lowest binding-site RMSD from each AChBP holo structure and number of clusters with representatives under 1, 1.5 and 2 Å RMSD obtained in the different simulations.

Clustering method		2BYQ	2BR7	2BYS	2PGZ	2BYR
				cMD		
Binding Site	Lowest (Å)	1.07	1.04	1.39	1.19	1.56
	# clusters < 1 Å	0	0	0	0	0
	# clusters < 1.5	3	2	1	3	0
	# clusters < 2 Å	5	4	3	8	5
Volumetric	Lowest (Å)	1.05	1.1	1.42	1.2	1.39
	# clusters < 1 Å	0	0	0	0	0
	# clusters < 1.5	4	4	1	1	3
	# clusters < 2 Å	4	5	5	7	6
				Single-Boost aMD		
Binding Site	Lowest (Å)	0.95	1.38	1.36	1.33	1.51
	# clusters < 1 Å	1	0	0	0	0
	# clusters < 1.5	2	1	1	1	0
	# clusters < 2 Å	6	4	4	3	3
Volumetric	Lowest (Å)	1.18	1.37	1.44	1.51	1.37
	# clusters < 1 Å	0	0	0	0	0
	# clusters < 1.5	2	2	1	0	1
	# clusters < 2 Å	7	6	5	4	3
				Dual-Boost aMD		
Binding Site	Lowest (Å)	1.52	1.23	1.67	1.90	1.59
	# clusters < 1 Å	0	0	0	0	0
	# clusters < 1.5	0	1	0	0	0
	# clusters < 2 Å	5	5	6	1	2
Volumetric	Lowest (Å)	1.33	1.41	1.56	1.35	2.02
	# clusters < 1 Å	0	0	0	0	0
	# clusters < 1.5	2	2	0	1	0
	# clusters < 2 Å	6	4	7	2	0

Both clustering methods identified representative conformations close to the holo structures in the cMD simulation, with slightly lower RMSDs for the ensemble produced by volumetric clustering.

The ensemble derived from single-boost aMD showed similar characteristics, while clustering of the dual-boost aMD trajectory gave higher RMSDs, as expected given the low percentage of frames close to the holo structures sampled in the whole trajectory. The ensemble of 10 cluster representatives obtained by volumetric clustering of the cMD trajectory spans different degrees of opening of the C-loop conformation (Figure 3.4), accurately representing the range of loop fluctuations identified in the available crystal structures. Therefore, we used the above ensemble of protein conformations to assess whether ensemble docking can reproduce the experimental ligand orientations in the five complexes of the AChBP.

We first evaluated the ability of the Glide XP docking method and the related computational protocol (see Par. 3.2 Methods) to reproduce the experimental binding geometries when ligands are docked to the X-ray holo protein structure (redocking). The results

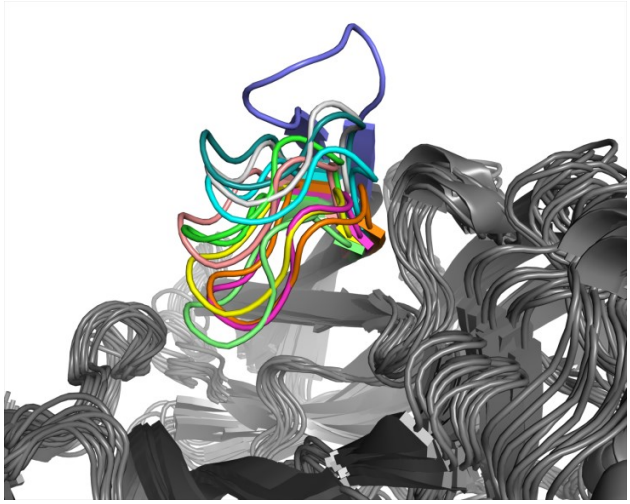

Figure 3.4: Ensemble of loop conformations in the 10 cluster representatives obtained with the volumetric clustering methods applied to the apo AChBP cMD trajectory.

are reported in Table App. A2. The correct geometric placement of the ligand was assessed by calculating the dRMSD between the ligand-site distances in the docking pose and the corresponding distances in the X-ray complex (see Par. 3.2 Methods). The best-scored pose was correctly placed in the binding site for all the ligands, except for COC, for which only the fourth scored pose was properly oriented. In all cases, the pose with the lowest dRMSD value gave a satisfactory reproduction of the experimental binding geometry (Figure App. A2). Then, we investigated the docking of each ligand to the conformational ensemble obtained from the apo protein simulation. Given that the binding-site conformations in the ensemble are not modeled around a ligand, they may not be appropriate for accepting ligands. Docking was performed using both the standard Glide XP protocol and by softening the van der Waals potential (see Par. 3.2 Methods). This latter soft-docking strategy (Soft XP) performed slightly better and produced poses with dRMSD < 2 Å to the X-ray complex for all the ligands (Figure 3.5a). Overall, the best geometric poses (lowest dRMSD values) gave a satisfactory reproduction of both the protein and the ligand experimental conformations (Figure 3.6). As expected, the dRMSD values were higher than those obtained in the redocking runs due to the differences between the side-chain conformations in the ensemble and in the reference holo structures.

Relying on the success of the proposed ensemble-docking approach, we assessed methodological choices that could be relevant to its effective use: the ability of the Glide Soft XP docking approach to correctly rank the best geometrical pose; the effect of increasing the protein conformational ensemble size on the pose definition and ranking; and the comparative performances of different docking methods.

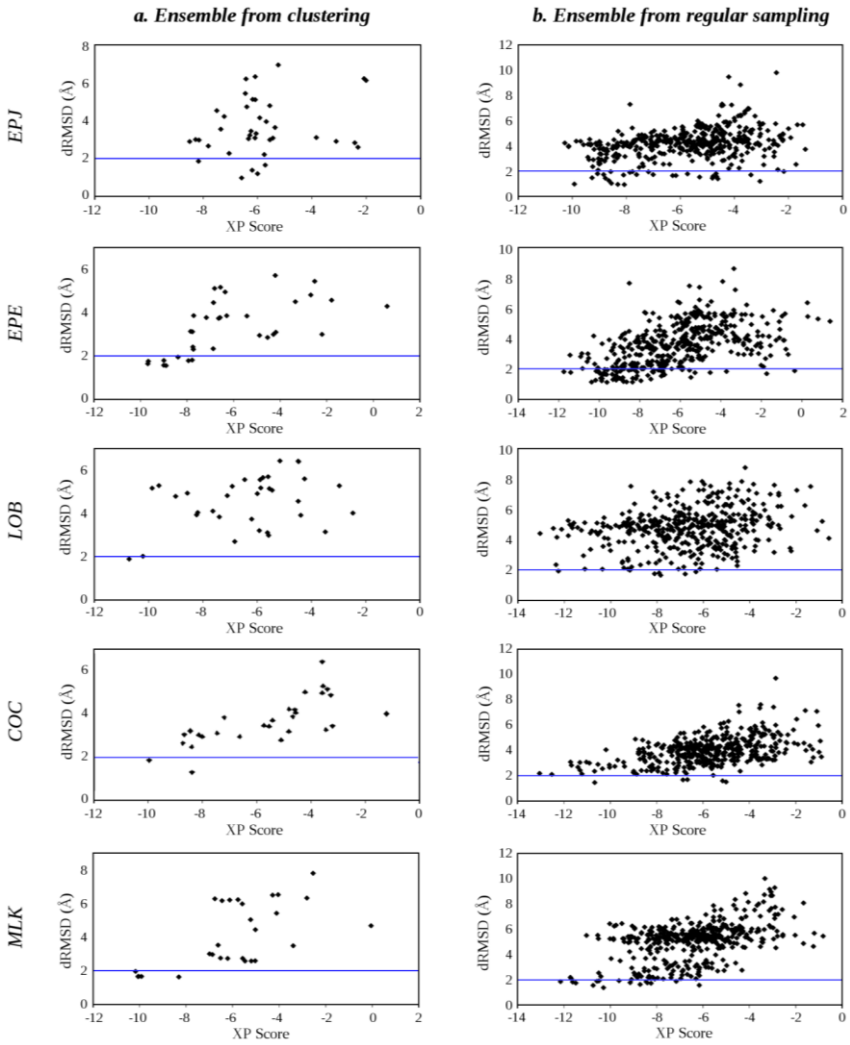

Figure 3.5: Results of ensemble docking to AChBP. The conformational ensemble was obtained by: a) volume clustering of the cMD trajectory; b) regular sampling of the cMD trajectory (every 2 ns). Plots show the dRMSD values to the X-ray geometry vs. the Glide XP Score.

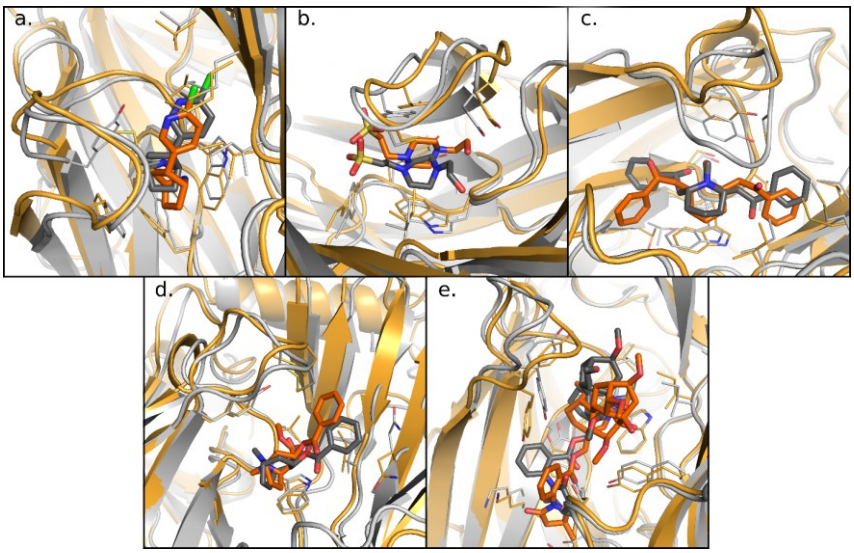

Figure 3.6: Docking poses with the lowest dRMSD to the AChBP X-ray structures obtained by ensemble docking (orange) superimposed on the X-ray binding structures (gray) for: a) EPJ; b) EPE; c) LOB; d) COC; e) MLK.

For the first task, we performed ensemble docking to the 10 representative protein conformations obtained by volume clustering; for 4 out of 5 ligands, the Glide XP scoring function successfully identified the poses closest to the X-ray complex (dRMSD values lower than 2 Å, Table 3.3a). The method penalized only the best dRMSD pose of the EPJ ligand, which is located in a small sub-cavity of the binding site that is often made inaccessible by protein side chains. Moreover, we analyzed the relationship between the binding-site RMSD to the X-ray holo structure of each conformation in the ensemble derived by volume clustering and its performance in ensemble-docking.

Table 3.3: Evaluation of the docking poses for the AChBP ligands: the dRMSD to the X-ray complex is reported for the best-scored poses (Best Score) and the best geometric poses (Best dRMSD) obtained with different docking methods and different ensemble-docking approaches.[a]

		dRMSD (Å)			
		Methods			
Ligand	Docking pose	(a) Soft XP Volume Clustering	(b) Soft XP Regular Sampling (2 ns)	(c) Soft SP Volume Clustering	(d) Rescored Soft SP Volume Clustering
EPJ	Best Score	2.89	4.22	1.7	2.3
	Best dRMSD	0.97	0.93	0.99	0.99
EPE	Best Score	1.63	1.81	1.9	1.69
	Best dRMSD	1.55	1.11	1.44	1.44
LOB	Best Score	1.88	5.24	3.74	4.88
	Best dRMSD	1.88	1.64	1.81	1.81
COC	Best Score	1.48	2.18	3.28	3.28
	Best dRMSD	1.48	1.46	1.67	1.67
MLK	Best Score	1.96	1.85	1.88	1.88
	Best dRMSD	1.65	1.37	1.3	1.3

[a]*Conformational ensemble obtained by volume clustering or by regular sampling*

In contrast to expectations, we found that the conformations closest to the experimental holo structure did not necessarily produce the best docking pose, and they often failed to correctly place the ligand in the binding site. This could be due to small deviations in one or a few side chains from the experimental conformations affecting the docking pose. This implies that a limited conformational ensemble is not sufficient to explain the whole combinatorial possibility of side-chain orientations. To determine whether increasing the variety of the ensemble could lead to better results, an ensemble-docking run on 500 protein conformations, extracted by regular sampling from the cMD trajectory, was performed for all the ligands using Glide Soft XP. The results were similar to those found with the conformational

ensemble derived by clustering (Figure 3.5b), with some poses showing lower dRMSD (probably due to the high number of generated poses). However, the scoring function had greater difficulty in identifying the best poses (Table 3.3b), probably due to an increase in false positives. Overall, the results achieved using an extended ensemble do not justify the extensive computational cost of 500 docking calculations per ligand.

Finally, we compared the performance of the Glide standard precision (SP) method, including the soft-docking approach (Soft SP), with that discussed above for the Soft XP. An alternative strategy that consists of rescoring the Glide Soft SP poses with the XP score (Rescored Soft SP) was also assessed. As shown in Table 3.3, all the methods tested generated poses close to the X-ray geometry ($0.9 <$ dRMSD < 1.9) and showed similar dRMSD values for the same ligand. However, both Soft SP and Rescored Soft SP produced a high number of incorrect best-scored poses (Table 3.3c,d). No particular advantage was introduced by XP rescoring.

Specific binding of D-allose to the allose binding protein

For our studies, we selected the X-ray depositions of the E. coli D-allose binding protein structures in the open/apo (PDB ID: 1GUD)[152] and closed/holo (PDB ID: 1RPJ)[151] conformations (Figure 3.7).

The structures consist of two similar Rossmann fold domains linked by a three-stranded hinge region responsible for the hinge-bending domain motion. The sugar-binding site is buried at the interface between the two domains. Strong binding with D-allose is enabled by a characteristic network of hydrogen bonds with residues in both protein domains and by stacking interactions with three aromatic rings forming a binding cleft that is perfectly designed for

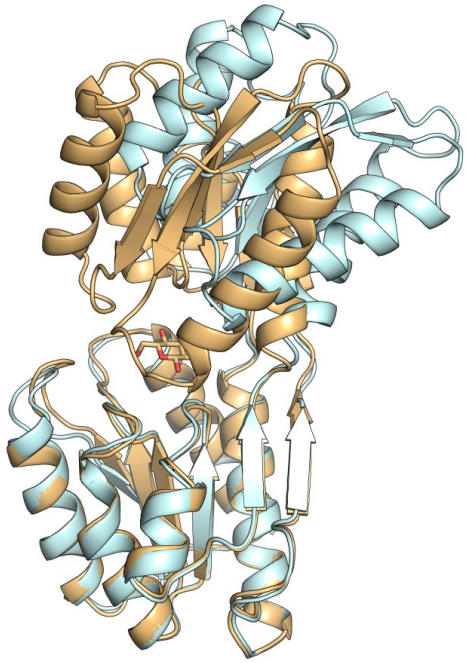

Figure 3.7: Structure of the open/apo (cyan, PDB ID: 1GUD) and closed/holo (orange, PDB ID: 1RPJ) X-ray conformations of the Allose BP. The allose molecule is shown as orange sticks.

binding this sugar molecule. Both of these features contribute to the high binding specificity of this receptor for D-allose.[151]

Also for this system we verified the possibility of reaching the protein holo conformation from MD simulations of the apo structure by adequately sampling its energy landscape. A slower transition between the open and closed states, associated with a high energy barrier, was expected. Therefore, this case was particularly suitable for evaluating the ability of aMD techniques to enhance sampling and reach conformations far from the starting structure. First, 1.0 μs of cMD simulation on the Allose BP open/apo structure was performed and compared to aMD simulations with four different sets of acceleration parameters (Table App. A1). The single-boost aMD was set with a boost on the potential energy slightly higher than the standard value to partially compensate for the lack of the dihedral

boost. This simulation was also extended over 1 μs to improve the probability of reaching the closed protein state. For the dual-boost simulations, three sets of parameters with increasing boost were used: "*dual-boost aMD 1*" with reduced acceleration parameters; "*dual-boost aMD 2*" with standard acceleration parameters (see Par. 3.2 Methods), and "*dual-boost aMD 3*" with increased acceleration parameters.

To analyse the effectiveness of sampling the hinge-bending motion of the two domains associated with the transition between the open and closed states of the Allose BP, the conformational state probabilities from the different simulations were projected in the subspace defined by two geometric variables: the angle that defines the degree of closure of the binding-site region and the dihedral angle describing the relative orientation of the two domains (Figure 3.8).

Despite the extended simulation time, the closed conformation was never approached in the cMD simulation; only fluctuations around the apo conformation and transitions to more open states were

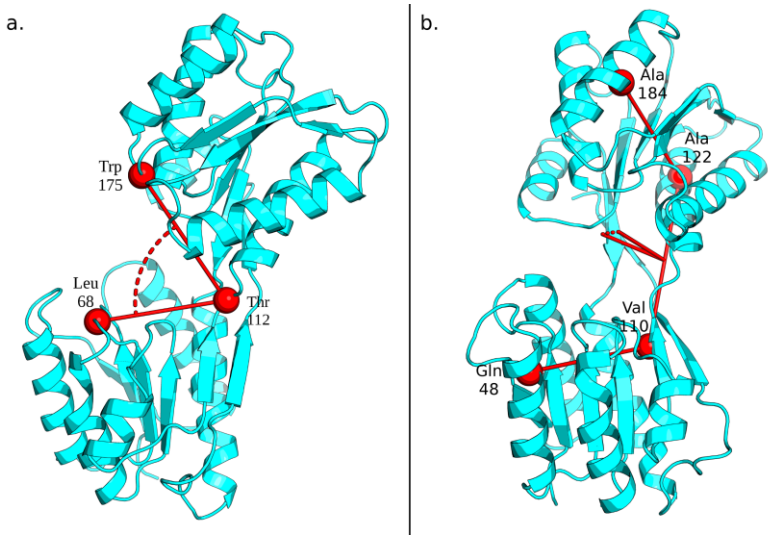

Figure 3.8: Variables selected to describe the sub-space related to the domain hinge-motion of the Allose BP during the simulation: a) angle defining the closure of the two domains; b) dihedral angle defining the mutual orientation of the two domains.

observed (see the top panel of Figure 3.9). Compared to the cMD, the single-boost aMD simulation in 1.0 μs explored a wider region of the conformational subspace defined by the two variables, approaching conformations near the closed state (Figure 3.9, central panel). This led us to extend this simulation to 1.55 μs (Figure 3.9, bottom panel), thus allowing sampling of the region around the closed/holo conformation.

Additional dual boost aMD simulations were performed with different sets of acceleration parameters. Dual boost aMD 2, with parameters calculated from the formula presented in Par. 3.2 Methods; dual boost aMD 1, with decreased boost parameters; dual boost aMD 3, with increased acceleration parameters. All the dual-

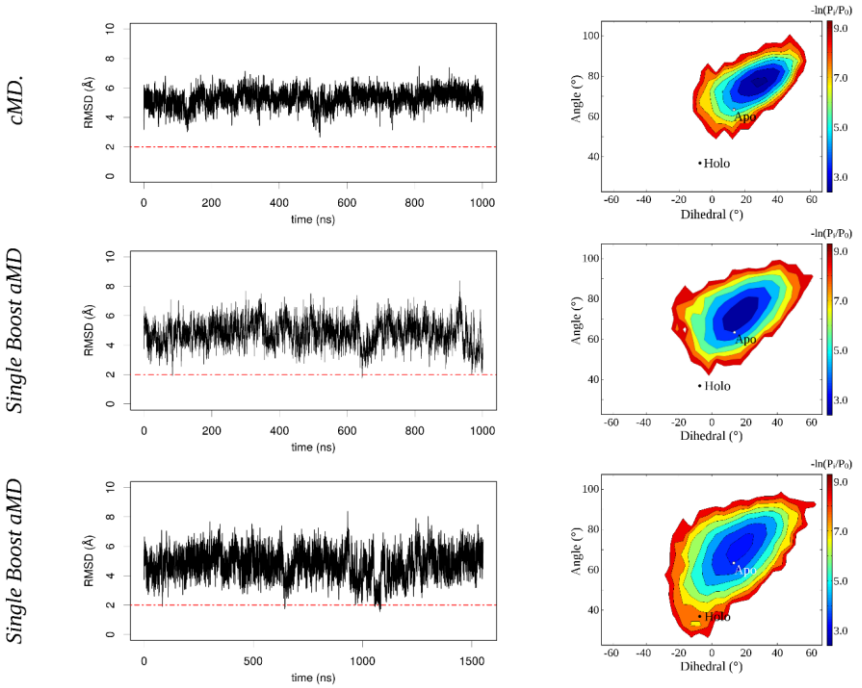

Figure 3.9: RMSD plots (left) and maps of the conformational state probability (right) for the Allose BP cMD (top panel), single-boost aMD at 1 μs (central panel) and 1.55 μs of simulation time (bottom panel). The RMSD from the holo X-ray structure is calculated on the heavy atoms of the binding-site residues. Probability maps are reported for the subspace defined in Figure 3.8. Apo and holo X-ray structures are projected onto the maps.

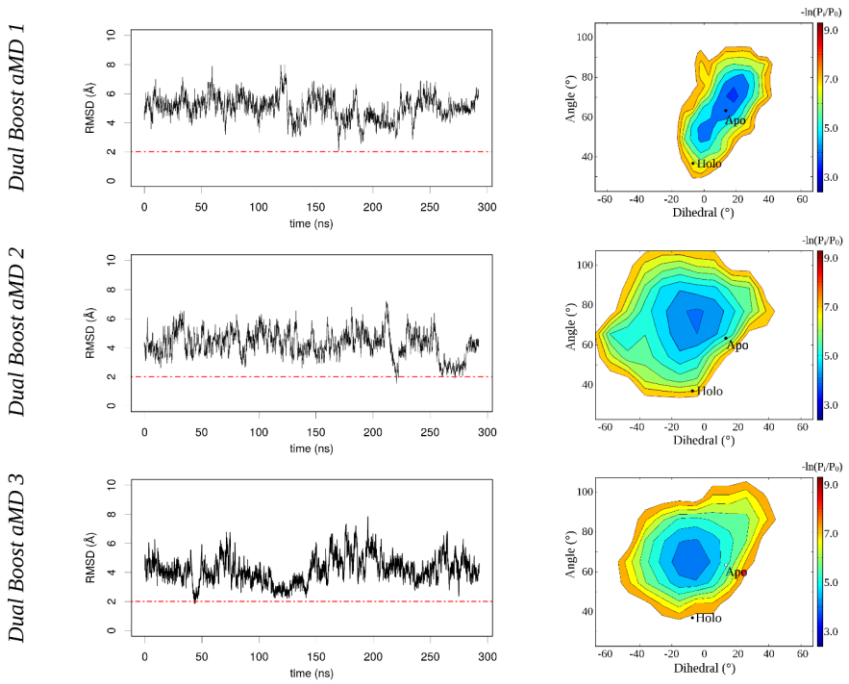

Figure 3.10: RMSD plots (left) and maps of the conformational state probability (right) for the Allose BP dual-boost aMD with different acceleration parameters. The RMSD is calculated from the holo X-ray structure on the heavy atoms of the binding-site residues. Probability maps are reported for the subspace defined in Figure 3.8. Apo and holo X-ray structures are projected onto the maps

boost aMD simulations (Figure 3.10) reached conformations near the closed one and allowed sampling of a wider space, especially along the dihedral coordinate, in a shorter time (300 ns) with respect to the single-boost simulations. However, progressively higher boosts resulted in increasingly homogeneous sampling in each direction, losing the original surface topology (Figure 3.10, central and bottom panels). This entails a lower percentage of conformations in the most frequently sampled zones of the obtained trajectories.

To evaluate the possibility of reproducing the experimental structure of the protein-ligand complex starting from an ensemble of representative conformations in the apo protein landscape, we used the single-boost aMD trajectory and applied clustering and ensemble

docking. Again, we evaluated the two clustering methods presented above, one based on the binding-site residue positions (GROMOS) using a cut-off distance of 2.0Å and the other based on the binding-site accessible volumes. In this case, the first method produced many clusters with representatives far from the holo structure (Table App. A3). This can be attributed to the long simulation time spent to explore open conformations with different binding sites. Only one cluster representing a semi-closed state (cluster 5) was found in the 10 most populated clusters, and only the fifteenth cluster could be considered a suitable candidate to represent the closed conformation.

Changing the target variable for cluster analysis led to a significant improvement (Table 3.4a). The clustering method based on similarity of the binding-site accessible volumes (see Par. 3.2 Methods) grouped all the open conformations into a single huge cluster (cluster 1) and obtained many clusters describing conformations with different binding-site volumes, some representing closed and semi-closed protein states (Figure 3.11).

Initially, the ability of docking to reproduce the experimental structure of the complex was evaluated by docking the allose molecule to the holo X-ray protein structure. Good redocking results were obtained with Glide XP, as shown by the very low dRMSD to the experimental binding geometry of the best pose (XP Score = -10.72; dRMSD = 0.31 Å). It is conceivable that the high specificity of the intermolecular interactions involved in ligand binding helped the docking to detect the correct ligand orientation.

Table 3.4: Results obtained by clustering the Allose BP single-boost aMD trajectory with the volumetric clustering method and applying ensemble docking to the cluster representatives.

a. Clustering			b. Docking					
Cluster	Cluster size	Binding site RMSD (Å)		Pose 1	Pose 2	Pose 3	Pose 4	Pose 5
1	90.66%	5.50	dRMSD (Å)	8.11	7.28	7.27	-	-
			XP Score	-5.22	-4.84	-4.25		
2	1.88%	1.93	dRMSD (Å)	1.58	1.51	1.46	1.42	2.06
			XP Score	-5.57	-5.46	-5.39	-5.39	-5.00
3	1.26%	3.06	dRMSD (Å)	4.21	4.26	3.57	3.59	3.57
			XP Score	-4.62	-4.07	-3.38	-3.69	-3.48
4	3.22%	1.75	dRMSD (Å)	3.26	3.51	2.57	3.50	-
			XP Score	-6.04	-5.76	-5.74	-5.66	
5	0.28%	1.95	dRMSD (Å)	2.99	2.96	2.99	2.57	-
			XP Score	-6.79	-5.92	-5.85	-4.83	
6	0.98%	2.24	dRMSD (Å)	4.03	3.79	-	-	-
			XP Score	-5.12	-4.48			
7	0.51%	2.19	dRMSD (Å)	3.20	3.20	3.18	-	-
			XP Score	-6.70	-6.53	-6.52		
8	0.35%	2.68	dRMSD (Å)	3.68	3.37	3.35	3.35	3.38
			XP Score	-5.74	-5.40	-5.39	-5.36	-5.35
9	0.78%	1.42	dRMSD (Å)	3.06	3.06	-	-	-
			XP Score	-5.74	-5.42			
10	0.08%	1.79	dRMSD (Å)	3.94	3.92	3.89	3.94	-
			XP Score	-5.57	-5.41	-5.32	-5.32	

The ability of the ensemble-docking strategy to reproduce the experimental binding geometry was tested on the ensemble of 10 representative conformations obtained from the volumetric clustering using Glide XP. Despite the high number of suitable representatives (5 out of 10 conformations with binding-site RMSD to the holo conformation < 2 Å), only that belonging to cluster 2 provided good

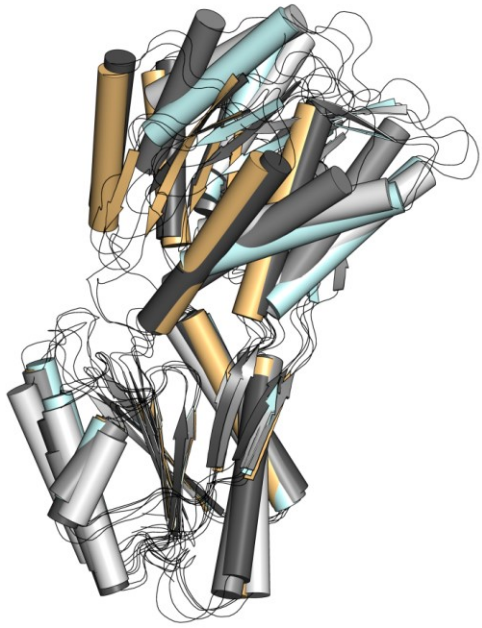

Figure 3.11: Three selected conformations spanning the hinge-bending motion of the Allose BP domains obtained by volumetric clustering (different shades of gray) are shown superimposed on the apo (pale-cyan) and holo (light-orange) X-ray protein structures.

docking results (Table 3.4b). Other components of the ensemble, with even lower RMSDs to the holo conformation, failed to correctly place allose in the binding site. This low success rate can be explained by the highly specific hydrogen bonding and aromatic stacking interactions that characterize this binding process. The correct orientation of a few side chains involved in such interactions is more important than the global geometric reproduction of all the binding-site side-chain conformations. Moreover, the geometrically correct binding pose showed a less favourable docking score (-5.57) than those of the incorrect poses and the re-docked pose (-10.72); this implies a lack of some key interactions. To assess the stability of this pose (the best-scored pose derived from cluster 2), we performed a 50 ns cMD simulation. In the first few ns of the simulation, the protein adapted its conformation to the presence of the ligand, with an evident induced-fit effect that also involved the backbone of the protein. The

average structure obtained from this simulation recovered some of the key interactions with the allose ligand. In particular, the aromatic stacking with Trp 175 and Phe 15 was restored after a few ns of simulation and other polar interactions, already present in the docked pose, were further optimized with the relaxation of the protein side chains. These changes produced a stable complex with a binding geometry similar to the X-ray holo structure (Figure 3.12).

Other MD runs, performed starting from incorrect binding poses, produced unstable simulations in which allose escaped from the binding site and the protein assumed an open conformation in a few ns.

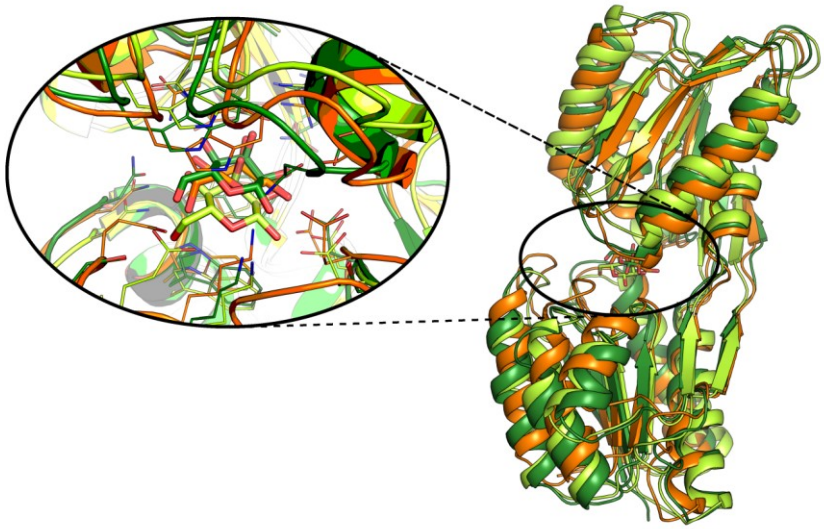

Figure 3.12: Best pose from Glide XP docking to the representative Allose BP conformation of cluster 2 (light-green) superimposed on the X-ray holo structure (orange). The average structure obtained from 50 ns of cMD starting from the docked pose is shown in dark-green.

3.4 Discussion

We report a comparison of two binding processes characterized by large conformational changes of a protein. We have identified the most appropriate methodological choices for predicting the ligand-binding geometries in the two mechanistic scenarios within the framework of the ensemble-docking approach.

In the binding mechanism of the AChBP system, the motion of a highly flexible loop governs the accessibility of multiple diverse ligands to the binding site, and molecular recognition is mainly driven by steric complementarity and weak dispersive interactions with the residues lining the pocket. Accordingly, our cMD simulations of the apo protein suggest the presence of a thermodynamic equilibrium among different conformational states separated by low energy barriers. These states included the experimentally determined AChBP bound structures, in which the C-loop adopts closed (with nicotinic agonists) or open (with antagonists or non-competitive ligands) geometries. Our results on the ensemble docking of multiple ligands to conformations representative of the apo protein energy landscape show that each ligand selects a suitable conformation, leading the system to a distinct state and accurately reproducing the known experimental holo structure. This result is in line with previous evidence of a conformational selection mechanism based on a virtual screening study using the relaxed complex scheme.[181]

The AChBP mechanism is an example of a type of binding processes often associated with local hinge-type motions at and around the binding site, characterized by the presence of several receptor conformations separated by low energy barriers. Conformational selection was proposed as the preferred mechanism of action for these types of processes.[182] Additionally, our findings support the general observation that binding events characterized by short-range dispersive interactions tend to favour population-shift

pathways.[8,12] Conversely, strong and long-range (ionic or dipole-dipole) or direct (hydrogen bond) ligand-protein interactions tend to favor the induced-fit mechanism.[8,12] Ligand binding to the Allose BP, associated with large hinge-bending motions of the two domains delimiting the binding site, is characterized by high binding specificity for D-allose due to strong and directed interactions and limited accessible space in the site. Our results suggest an interplay of conformational selection and induced fit. In fact, the aMD simulations demonstrated that closed states exist in the apo protein conformational landscape and can be sampled with an appropriate setting of the boost and simulation time. Docking of D-allose to one of these conformations, followed by a short MD relaxation of the docked pose, confirmed that further conformational changes are induced by the ligand and help to reach the experimental bound state. Several studies on other periplasmic binding proteins have focused on their large conformational change from an open to a closed state in the presence of a ligand, known as the "Venus flytrap mechanism".[183] Experimental and computational analyses of the maltose binding protein[184–186] indicated the existence of a dynamic equilibrium between a major open and a minor semi-closed conformational state in the unbound protein, assisted by solvation effects and packing of non-polar side chains. On this basis, a two-step mechanism was hypothesized for ligand binding, involving a population shift followed by induced fit to reach the fully closed state.[184] An initial conformational selection mechanism related to global protein motion with a local induced fit completion of the binding event were observed in several other systems. For example, local induced-fit adjustments near the binding site after initial conformational selection within the rest of the protein are a significant component in the binding of ubiquitin.[8,14]

Our studies provide clear indication that ensemble-docking technique can produce reliable predictions of the structure of ligand-

protein complexes in both mechanistic scenarios when suitable methodological choices are made.

First, enhanced sampling methods can play a key role in generating suitable structural ensembles, including the relevant large conformational changes. Accelerated MD appears particularly suitable given that it does not rely on the advance definition of the reaction coordinate. Here, we show that when the protein dynamics are characterized by transitions among conformational states separated by low energy barriers on a fast timescale (the AChBP system), conventional MD simulations are already effective in extensively sampling the energy landscape, particularly near the holo minima. In these cases, aMD may enhance the sampling of higher-energy regions, reducing the relative occurrence of some of the lowest-energy states relevant for binding. This in turn can reduce the chance of selecting relevant representative conformations by geometric clustering for the subsequent ensemble docking. In contrast, when transitions between the unbound and the bound states are on a slower timescale and associated with higher energy barriers (e.g., Allose BP), cMD has limited applicability, and aMD appears particularly suitable to enhance the sampling and reach conformations near the holo state. However, in these cases, particular attention has to be given to both the required boost and the simulation time. We have observed that dual-boost simulations with excessive boost can result in a wider achievable sampling space, which must be supported by an increased simulation time. Therefore, the aMD parameters should be set carefully to achieve the correct balance between wide sampling of the energy landscape and the time spent in regions relevant to the process.

Second, ensemble docking also requires an appropriate strategy to select a reduced set of conformations that are relevant to binding from the large collection in the MD trajectory. Several clustering methods have been proposed with this aim, but the most desirable strategy

would be to use kinetic clustering, where clusters are directly related to the metastable states of the underlying free-energy landscape.[187] In most applications, structures can only be clustered by geometrical similarity.[146,147] Here, we propose geometric clustering of the binding site to select the protein conformations relevant to describing the binding process. Two strategies were presented and assessed: one based on clustering the positions of the binding-site heavy atoms, and one specifically developed to improve the description of different cavity shapes based on the overlap of accessible volumes. Both methods performed well in the case of AChBP, providing a small set of conformations spanning the different degrees of opening of the C-loop, some very close to the experimental holo structures. The ensemble-docking results of this system also indicate that the selection of a cluster representative with low binding-site RMSD to the holo structure does not guarantee a successful reproduction of the binding geometry by docking because small deviations in a few side chains from the experimental conformations may affect the docking pose. This suggests the use of a larger number of conformations to include as many of the different side-chain orientations as possible. However, this choice is known to increase the false positive rate.[25] Alternatively, an additional criterion to select a subset of relevant conformations could be applied. For example, it has been proposed to perform the selection based on the correlation between the experimental binding affinities and the docking scores on a small number of known active compounds.[51,188] In our studies on the Allose BP, volumetric clustering of the binding site outperforms clustering of the positions of the heavy atoms, and it is able to identify a variety of clusters representing different binding-site volumes, including some related to closed and semi-closed protein conformations. However, only one cluster representative leads to a binding pose in agreement with the experimental structure. Again, this indicates the need for an accurate reproduction of the specific side-chain conformations, particularly when highly specific interactions drive the binding.

Third, the choice of the docking procedure deserves some attention. The AChBP and the Allose BP studies demonstrate that the Glide method, with adequate preparation of the ensemble of receptor conformations, can generate poses similar to the holo crystallographic structures. The test performed for binding to the AChBP indicates that this result is achievable with both the Glide SP and XP protocols. Although the scoring functions are not always successful in assigning the highest score to the pose closest to the crystal conformation,[189,190] in this study, we obtain a high success rate (4 out of 5 ligands with the top-scored pose with very low dRMSD to the X-ray complex) using Glide XP. The better performance of the XP protocol may be attributed to its more extensive sampling and to a scoring function with greater requirements for ligand-receptor complementarity than that of the SP protocol.[174] These characteristics have proven to be effective in screening out false positives[174] and are particularly important in our approach, where the ligand is docked to multiple receptor conformations. Regarding the need for a refinement stage for the docked poses, we observe that in ensemble docking of AChBP, driven by weak dispersive interactions, the soft-docking technique is sufficient to introduce some degree of local flexibility, which is useful for reproduction of the holo structures. Conversely, in the Allose BP binding, characterized by strong and directional interactions, short MD simulation on the docked pose helps to recover the key interactions observed in the experimental holo structure, suggesting the importance of a post-docking refinement stage for such mechanisms.

While obtaining reliable predictions of the structure of ligand-protein complexes in binding processes with large conformational changes remains a challenging task, this work, along with many others, provides clear indications that ensemble docking is an effective technique to handle protein flexibility in such studies. Moreover, our results highlight that knowledge of both the apo protein dynamics and the ligand-protein interactions involved in binding provides the

basis to develop the most appropriate methodology for successful ensemble-docking applications. Further studies focused on a wider set of ligand-binding processes are needed to confirm these encouraging findings and to develop appropriate protocols for virtual screening applications.

"Almost all aspects of life are engineered at the molecular level, and without understanding molecules we can only have a very sketchy understanding of life itself."

Francis Crick,
What Mad Pursuit: A Personal View of Scientific Discovery (1988)

CHAPTER 4

MOLECULAR DYNAMICS OF HIF-2A:ARNT LIGAND-INDUCED INHIBITION

4.1 Introduction

Hypoxia inducible factors (HIFs) are obligate heterodimers belonging to the basic helix–loop–helix (bHLH) superfamily of transcription factors that mediate the physiological responses to hypoxia. This extensive protein family is characterized by a 4–6 basic amino acids next to a HLH dimerization domain, both required to

properly bind DNA targets. Within the bHLH superfamily HIFs belong to the subfamily containing the PER/aryl hydrocarbon receptor nuclear translocator (ARNT)/single minded (SIM) (PAS) homology domain (bHLH-PAS).[191–193] Based on their heterodimerization behaviour, bHLH-PAS proteins can be further divided into two classes: class I members only form heterodimers with a member of class II, which, by contrast, can promiscuously homo- and heterodimerize. Class I includes aryl hydrocarbon receptor (AhR), aryl hydrocarbon receptor repressor (AhRR), single minded proteins (SIM1 and SIM2), circadian locomotor output cycles kaput (CLOCK), neuronal PAS domain protein (NPAS) 1-4, and three HIF-α subunits isoforms, HIF-1α, HIF-2α, and HIF-3α, each targeting both shared and distinct genes.[194] When transcriptionally active, HIF-α subunit dimerizes with the constitutive ARNT (also known as HIF-β) subunit, the best characterized class II protein; other members of this class include the tissue restricted ARNT2, and the circadian rhythm proteins BMAL1 and BMAL2.[191,193]

The C-terminal region of bHLH-PAS proteins is highly variable in length and composition and hosts the transactivation domains (TAD) where the transcriptional coactivators are recruited to initiate the transcription.[192] By contrast, the N-terminal portion contains three well-defined domains: bHLH, PAS-A, and PAS-B. The bHLH domain offers the primary dimerization interfaces and, together with the protein partner, determines the target gene recognition.[195] Despite low sequence identity, the PAS domains show conserved three-dimensional structures in a wide range of prokaryotic and eukaryotic proteins.[191] They contribute to the dimerization and increase the specificity of partner choice.[196,197] PAS-A, in particular, prevents dimerization with non- PAS-containing bHLH proteins and participate to the binding of DNA sequences that differ from the prototypical E-box motif.[195] The PAS-B domain commonly functions as a signalling domain and can host hydrophobic cavities for small molecules and/or cofactors that relay environmental or metabolic

signals;[197] consequently to the binding, allosteric changes occur that affect the affinity for partner molecules.[198]

Oxygen concentration is monitored by the hypoxic response pathway regulated by HIFs. The oxygen-sensitive regulatory subunits HIF-1α and HIF-2α contain an oxygen dependent degradation domain (ODD) and a N-terminal TAD in addition to the prototypical bHLH, PAS-A, PAS-B, and C-terminal TAD. Under normoxia (20% O2), HIF-α is rapidly degraded by the ubiquitin-proteasome system.[192] In hypoxic conditions, HIF-α escapes degradation and, after translocation to the nucleus, heterodimerizes with ARNT, binds hypoxia response elements (HREs) in the enhancer regions of target genes, interacts with CBP-p300 complex and initiates transcription.[192] 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.[192]

In tumour 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.[199] HIFs are commonly upregulated in a broad range of cancers[200–202] where they contribute also to resistance to oxidative stress, epithelial–mesenchymal transition (EMT), and tumour invasiveness. HIF-1α and HIF-2α accumulation can also be caused by reduced degradation, as in Von Hippel-Lindau (VHL) syndrome, an inherited familial cancer syndrome where mutation of VHL causes its inactivation.[203] HIFs function can be affected by mutations, that have been observed in different carcinomas, brain gliomas, and skin melanomas.[204] 19 of these mutations are located in the bHLH-PAS-A-PAS-B segments: 4 engage the DNA binding domain, while all the others are either at the interfaces between HIF-α and ARNT, highlighting the key role of protein-protein interactions in the

stability and activity of HIFs, or in the hydrophobic pockets of PAS domains.[85]

Internal hydrophobic cavities are observed in all available structures of bHLH-PAS family within both their PAS-A and PAS-B domains.[205] It has been shown that AhR uses its PAS-B internal cavity for binding a diverse set of small molecules thus activating nuclear translocation, dimerization with ARNT and DNA binding.[206,207] More in general, it has been shown that ligand binding in the pockets of other PAS domains induces long distance conformational changes that affect protein-protein interactions,[208] suggesting that PAS cavities can contain allosteric sites.[197,209] As HIF-α:ARNT dimerization is essential to bind DNA and initiate transcription, destabilizing protein–protein interactions in this system represents an optimal therapeutic approach for tumour treatment. However, direct antagonizing of the interfaces with small molecules is pharmacologically demanding and often unsuccessful due to the troublesome identification of the key residues to target.[210] By contrast, exploiting PAS internal cavities offers potential advantages, especially in terms of selectivity. In addition to the PAS domain cavities, HIF-1α:ARNT and HIF-2α:ARNT structures have a pocket at HIF-α PAS-B:ARNT PAS-A interface, which has been targeted by acriflavin,[85] a molecule that acts as a potent inhibitor of both HIF-1α and HIF-2α dimerization with ARNT. Acriflavin is a mixture of trypaflavin and proflavin which bind the pocket with comparable affinities causing the destabilization of HIFα:ARNT heterodimers in cells.[211]

By targeting distinct genes,[194] HIF-1α and HIF-2α affect tumour progression activating different signalling pathways.[201] This finding highlighted the need of developing isoform-specific drugs. HIF-2α PAS-B domain contains a relatively large (290 $Å^3$) cavity that can be occupied by either water or small molecules with sub-micromolar affinities. These small binders have been shown to impair

heterodimerization of isolated PAS-B domains in vitro.[82,84] In the framework of extensive efforts directed to identify inhibitors of HIF-2α:ARNT dimerization,[212,83] a molecule has been recently developed, 0X3 (N-(3-chloro-5-fluorophenyl)-4-nitro-2,1,3-benzoxadiazol-5-amine), that is able to disrupt heterodimerization also in living cells.[213] The compound fails to bind to HIF-1α, which has a smaller cavity in PAS-B domain. Notably, some tumour associated mutations at HIF-2α PAS-B domains specifically alter residues that are in contact with 0X3.

Albeit the molecular details of 0X3 interaction with HIF-2α PAS-B have been unveiled, how the ligand binding destabilizes the HIF-2α:ARNT complex remains unexplained. The recently determined structures of the entire bHLH-PAS region of the HIF-2α:ARNT dimer in the unbound, DNA-bound, and inhibitor-bound (0X3 and proflavin ligands) forms[85] provide a sound basis for assessing the inhibition mechanism. A schematic view of the dimer assembly is reported in Figure 4.1, in which the two bHLH domains are linked in a pseudo-symmetric arrangement and the PAS domains interact asymmetrically. Besides the interfaces between corresponding PAS domains, there are also interfaces formed by HIF-2α PAS-B-ARNT PAS-A and HIF-2α intramolecular interactions between PAS-A and PAS-B and between PAS-A and bHLH domains. The lack of physical interaction between ARNT domains facilitates flexibility for arrangements with different partners. Indeed in the NPAS1: and NPAS3:ARNT heterodimers, ARNT PAS-B domain is slightly displaced in comparison with HIF-α:ARNT complex.[205] PAS-B cavity residues facing 0X3 in the HIF-2α:ARNT-0X3 complex are not significantly perturbed, while the PAS domains slightly shift one respect to the others, with major rearrangements occurring at the interface between the HIF-2α and ARNT PAS-B domains.[85]

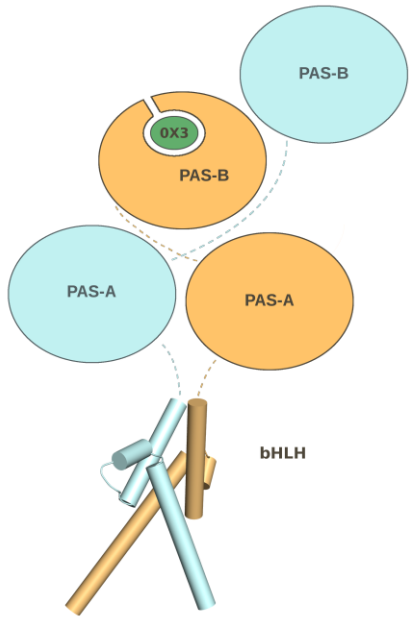

Figure 4.1: Simplified representation of the dimer assembly that illustrates the inter-domain interactions: ARNT domains in cyan, HIF-2α domains in orange, 0X3 ligand in green.

MD simulations provide valuable tools for the study of biomolecular interactions and for elucidation of the mechanisms by which substrate- or inhibitor-binding can alter the dynamics and function of a protein system (see Par. 1.1 Protein dynamics in biomolecular interactions and functions). To this aim, perturbations of protein dynamics (e.g. local dynamic fluctuations and secondary structure arrangements) upon binding can be analysed. Moreover, variations occurred in the free-energy can be investigated with different methods. Among these, MM-GBSA[214] has been successfully used both to characterize global free-energy differences and to highlight the hotspots for complex stabilization through per-residue decomposition analysis, providing results in good agreement with experimental mutagenesis data.[215–217] If the induced perturbations involve regions far away from the ligand binding sites, several approaches (see Par 1.1 Protein dynamics in biomolecular interactions and functions), including methods based on the study of correlated

motions in MD simulations,[64,66,67] can be used to detect residues that are allosterically coupled and pathways connecting the source of the perturbation with the effect (see Par. 1.1 Protein dynamics in biomolecular interactions and functions).

In this study we hypothesized that the ligand-induced local perturbation at the HIF-2α PAS-B domain dynamically propagates through the HIF-2α:ARNT dimerization interfaces by an allosteric inhibition mechanism. To study the functional dynamics of the complex and shed light into the mechanism of regulation of dimer stability, we compared the evolutionary, dynamical and energetic properties of HIF-2α:ARNT dimer structure in its unbound and 0X3-bound form. We identified both the molecular features of the ligand-induced perturbation and the key residues involved in inter-domain communication paths. This novel insight in HIF-2α regulation will guide the development of new specific inhibitors of aberrant HIF-2α activity.

4.2 Methods

System Preparation and Molecular Dynamics Simulations

Crystal structures for HIF-2α:ARNT dimer in its apo (PDB ID: 4ZP4) and holo (PDB ID: 4ZQD) forms[85] were obtained from the Protein Data Bank (PDB).[154] Unresolved regions in the apo deposition were modelled using the Rosetta all-atom de novo loop modelling method with the Next Generation Kinematic closure (NGK) procedure, a variant of the KInematic Closure (KIC) approach.[217–219] A starting set of 1000 loop models was generated with the parameters proposed by Conchúir and coworkers,[220] enabling the Taboo sampling feature and using Monte-Carlo simulated annealing

for rotamer-based side-chain optimization in a neighborhood of 10 Å around the loop structures. The ensemble of models was then clustered by backbone structural similarity using the Self Organizing Map (SOM) approach previously described.[147,221,222] The best conformation for each loop was selected from the most populated cluster by Rosetta energy score. Missing regions in the holo structure were completed by grafting the atomic coordinates of the loops modelled for the apo form and refined using Modeller 9v8.[223] The completed structures were then pre-processed for simulation with the Schrodinger's Protein Preparation Wizard tool[155]: hydrogen atoms were added, all water molecules were removed, C and N terminal capping were added, disulfide bonds were assigned and residue protonation states were determined by PROPKA[156] at pH = 7.0. Each system was then solvated in an octahedral box with about 59000 TIP3P water molecules, and neutralized with Na+ ions using the GROMACS[224] preparation tools. The total number of atoms of the system is 187832. The minimal distance between the protein and the box boundaries was set to 12 Å. Simulations were run using GROMACS 5.1[224] with Amber ff99sb*-ILDNP force-field[225]. 0X3 inhibitor in the holo structure was parameterised using GAFF.[162] 0X3 charges were calculated with the restricted electrostatic potential (RESP) method[92] at HF/6-31G* after ab-initio optimization of the ligand. A multistage equilibration protocol (modified from the one used by Fornili and coworkers[226]) was applied to all simulations to remove unfavourable contacts and provide a reliable starting point for the production runs: the system was first subjected to 1000 step of steepest descent energy minimization, followed by 1000 step of conjugate gradient with positional restraints (2000 kJ mol^{-1} nm^{-2}) on all resolved atoms. This minimization process was then repeated with weaker (1000 kJ mol^{-1} nm^{-2}) restraints on the backbone of resolved regions. Subsequently a 200 ps NVT MD simulation was used to heat the system from 0 to 100 K with restraints lowered to 400 kJ mol^{-1} nm^{-2} and then the system was heated up to 300 K in 400 ps during a

NPT simulation with further lowered restraint (200 kJ mol^{-1} nm^{-2}). Finally, the system was equilibrated during a NPT simulation for 1 ns with backbone restraints lowered to 50 kJ mol^{-1} nm^{-2}. All the restraints were removed for the production runs at 300 K, performed in three replicas of 300 ns each. In the NVT simulations temperature was controlled by the Berendsen thermostat,[103] while in the NPT simulations the V-rescale thermostat[104] was used with a time constant of 0.1 ps and pressure was set to 1 bar by the Parrinello-Rahman barostat[227] with time constant of 2 ps. A time step of 2.0 fs was used, together with the LINCS[99] algorithm to constrain all the bonds. The particle mesh Ewald method[102] was used to treat the long-range electrostatic interactions with the cutoff distances set at 12 Å.

Analysis of MD Simulations

Global structural changes during the simulations were monitored by RMSD. Average per-residue flexibility was measured by RMSF of the atomic positions. RMSD and RMSF values were calculated for the protein Cα atoms using the R[228] Bio3D package.[178] All RMSF values were computed on a trajectory obtained concatenating the three replicas, excluding the first 50 ns of each simulation. Secondary structure attribution was done with DSSP.[229] Cluster analysis of the inhibitor geometries in the binding pocket was performed using the GROMOS nearest neighbour algorithm[168] implemented in GROMACS analysis tools, after fitting on the Cα atoms of HIF-2α PAS-B domain.

The binding free energy of dimer formation was estimated using the Molecular Mechanics Generalized Born Surface Area (MM-GBSA) method[230,231], implemented in the AMBER software package[157,232]. In this method, the $\Delta G_{binding}$ is obtained as the sum of energy associated with complex formation in the gas-phase and the difference in solvation free energies between the complex and the

unbound monomers. The method includes an implicit solvent model. The polar solvation term was approximated with the Generalized Born (GB) model[233] using OBC re-scaling of the effective Born radii[234]. The non-polar solvation term was calculated as the product of the surface tension parameter and the solvent accessible surface area (SA) evaluated using the Linear Combination of Pairwise Overlap (LCPO) algorithm.[235] The single-trajectory approach[214] was used. In this approach both monomer conformations for the calculation were obtained from the dimerized state MD simulation instead of performing distinct simulations of the three different states (monomeric ARNT, monomeric HIF-2α, and bound state). MM-GBSA calculations were performed on a subset of conformations from the equilibrated portion of the MD simulations. For this purpose, the domain contributions were calculated on the last 20 ns of each replica and each energy component was determined by averaging over the contributions from all the conformers. Single interfaces were analysed using a per-residue energy decomposition. For this purpose, a common ensemble of conformations sampled in all three replicas was identified in the principal component subspace of inter-domain motions, calculated on the subset of Cα at domain interface (Figure 4.2).

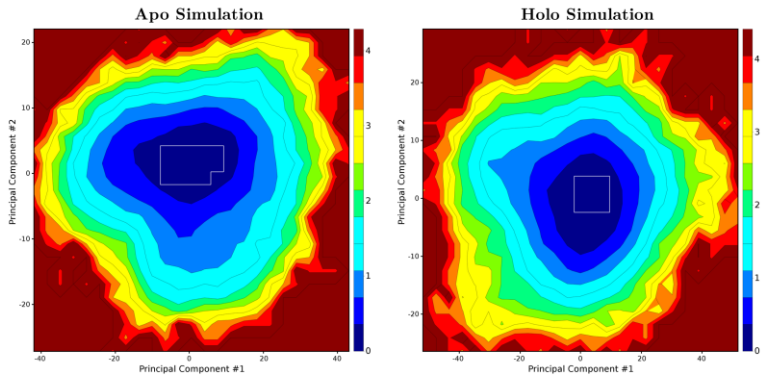

Figure 4.2: Probability density maps of conformations from the combined MD trajectories. Bins are calculated on the subspace of the first two principal component of motions for HIF-2α PAS-B and ARNT PAS-B. Apo (left panel) and holo (right panel) simulations. The white box contains the most populated bins, that include about 15% of the whole trajectory.

In this analysis, a residue of a domain A was considered within the A-B interface if at least one of its atoms was found within 3.5 Å from an atom of the B domain in at least 10% of the simulation. To identify protein segments with correlated atomic motions, a correlation network analysis[64] was performed using Bio3D[178],[179]. In this approach the cross-correlation coefficient was calculated from the displacement of all the Cα atom pairs.[236] A weighted graph was generated from the cross-correlation matrix, in which each residue represents a node and an edge is drawn when the absolute correlation between two residues is greater than 0.4. Edges with positive weights connect residues with correlated motion, while negative weights describe anti-correlated motions. Shortest and suboptimal path analysis,[179] conducted on the 50 shortest detectable paths, was used to highlight differences in inter-domain communication in the apo and holo states.

ConSurf Analysis

Residue conservation on protein surfaces was analysed with ConSurf.[237,238] PAS-domain sequences were detected with a PSI-BLAST[239] search (3 iterations; E-value cutoff 0.0001) of the PDB sequence of HIF-2α:ARNT (PDB ID: 4ZP4) against the UniProt database.[240] Orthologous sequences were manually selected for each protein independently: HIF-1α, HIF-2α and HIF-3α for HIF-2α, and ARNT1 and ARNT2 for ARNT, for a total of 110 and 170 sequences. Input multiple sequence alignments were generated with Muscle.[241]

4.3 Results

In the following, we present an analysis of: 1) the HIF-2α:ARNT dimer structure; 2) the intrinsic dynamics of the unbound dimer; and 3) the effect of 0X3 inhibitor on the dimer stability.

The domain structure of HIF-2α:ARNT dimer

Among the recently crystallised structures of the N-terminal bHLH-PAS region of the HIF-2α:ARNT dimer[85], we selected and simulated the unbound form (PDB ID: 4ZP4) and the inhibitor bound form with 0X3 ligand in the HIF-2α PAS-B domain (PDB ID: 4ZQD). Domain and secondary structure definitions for the two protomers are reported in Figure 4.3.

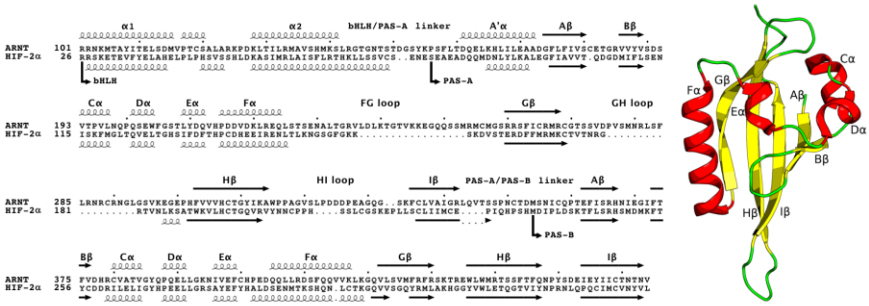

Figure 4.3: Domains and secondary structure elements of ARNT and HIF-2α in the bHLH-PAS region. On the left: the sequence alignment was obtained by Clustal Omega[242], secondary structure information was extracted by DSSP from the PDB file 4ZP4, monomer A and B for ARNT and HIF-2α respectively. Helices are displayed as squiggles and β-strands as arrows, and labelled according to the PAS domain nomenclature[209]. The figure was generated with the ESPript server[243]. On the right: a general PAS domain fold, with secondary structures labelled.

The dimer structure of each protomer includes three domains (bHLH, PAS-A and PAS-B), with only few unresolved segments encompassing the inter-domain (bHLH/PAS-A and PAS-A/PAS-B) linkers and the three PAS-A loops on each partner. Among these 10 segments, the GH loop and the PAS-A/PAS-B linker of HIF-2α are resolved in at least one of the available crystal structures. The other 8 elements were modelled using an extended Rosetta loop modelling protocol previously adopted in the modelling of AhR:ARNT dimer[217] (see Par. 4.2 Methods). An overall view of the dimer structure, including the modelled loops, is shown in Figure 4.4: the dimer has a compact core region formed by ARNT-PAS-A, HIF-2α-PAS-A and

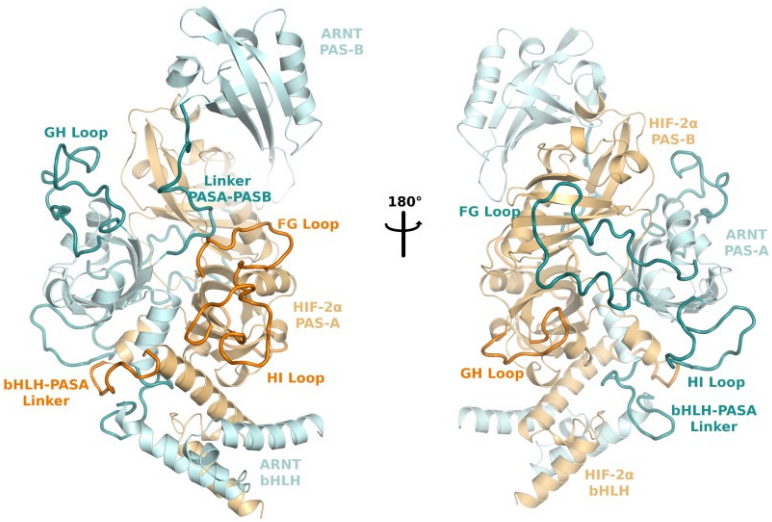

Figure 4.4: Structure of HIF-2α:ARNT dimer with the modelled loops and linkers. Cartoon representation of the unbound dimer structure (4ZP4): ARNT in cyan, HIF-2α in orange. Modelled segments are represented in darker colours.

HIF-2α-PAS-B domains; this is connected to the bHLH region on one side and the ARNT-PAS-B domain on the opposite side.

Mapping of residue evolutionary conservation on protein structure can provide reliable prediction of functionally relevant elements and their role in multi domain organisation.[244] Therefore, to investigate the relative importance of each loop and domain in the context of the dimer, we calculated the evolutionary conservation profile using the ConSurf Server Database.[237,238] Input sequences were selected from HIF-2α and ARNT orthologues and predictions were run independently for the two proteins to discriminate the specific role played by each partner.

The conservation profiles highlight highly conserved patches on the bHLH and PAS core domains, especially for the residues lying at the dimerization interfaces (Figure 4.5 and annotated sequences in Figure App. B1). As expected, the most conserved region is the bHLH portion responsible for DNA binding, while loops are generally poorly conserved.

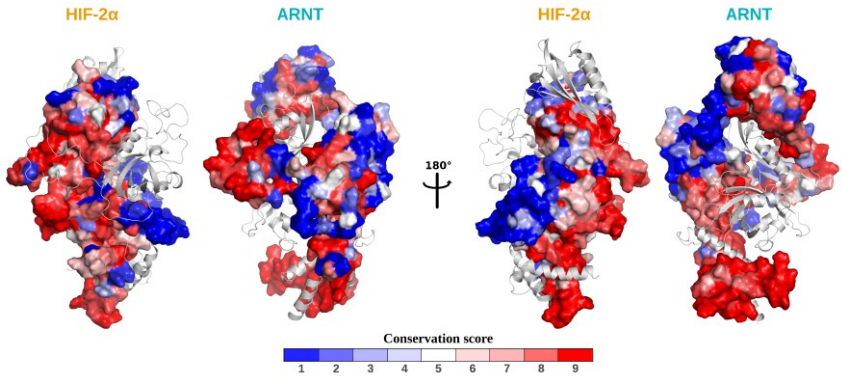

Figure 4.5: Residue evolutionary conservation mapped on the solvent accessible surfaces. In each representation, the solvent accessible surface is shown for one protein, while the protein partner is represented in light-grey cartoon. Evolutionary conservation scores obtained by ConSurf (1 poorly conserved, 9 highly conserved) are reported in a blue-white-red colour scale.

It should be noted that most of the modelled loops belonging to the PAS-A domains resemble structural embellishment with a typical Ω-loop shape and no expected functional role, except for: the ARNT-PAS-A FG loop, the HIF-2α-PAS-A GH loop and the HIF-2α-PAS-A/PAS-B linker. The last two are resolved in the X-ray structure and have a functional role: the HIF-2α-PAS-A GH loop is known to bind DNA at a distance six base pairs away from the hexameric core element;[85] the HIF-2α-PAS-A/PAS-B linker is buried, and lies at the ARNT-PAS-A:HIF-2α-PAS-B interface therefore suggesting a role in the dimerization. No information about the ARNT-PAS-A FG loop functional role has been reported. In HIF-2α, the PAS-B C-terminal linker, including a loop and a short α helix and inserted into the PAS-B:PAS-B interface, shows some conserved residues (Figure App. B1) thus suggesting a putative role in the dimerization.

Analysis of HIF-2α:ARNT dynamics

The dynamics of the unbound HIF-2α:ARNT dimer (PDB ID: 4ZP4) was investigated to highlight the role of domains and secondary structure elements in the dimer stability as well as to characterise the flexibility of the inter-domain interfaces. A set of three replicas of 300 ns MD simulations were performed. The RMSD plots of the core domains (Figure App. B2) show well equilibrated trajectories after 50 ns. The linkers and loops were removed from the calculation of the RMSD due to the expected high flexibility of these regions. Indeed, high flexibility in the PAS-A loops is also evident from the root mean square fluctuation (RMSF) plot of the concatenated trajectories for the complete system (Figure App. B3), while the bHLH regions show enhanced flexibility due to their terminal position and lack of DNA interactions. As shown in the previous subsection (see The domain structure of HIF-2α:ARNT dimer) ARNT PAS-B domain is involved in fewer interactions with the rest of the protein. Consistently, in all three MD simulations we detected a reorientation of ARNT PAS-B (Figure App. B4), which seems to be an intrinsic feature of this monomer and could be instrumental to select suitable conformations to bind different partners, as recently suggested from a deposition of NPAS1:ARNT dimer.[205]

To extract the profile of local intra-domain flexibility, we produced separated RMSF plots of each PAS domain after concatenation of the three trajectories and independent frame superposition of the Cα atoms of each domain (Figure 4.6). A' helices of both PAS-A domains show high flexibility, especially in ARNT, while other secondary structure elements are generally rigid, with relatively high RMSF values only in correspondence of connective loops. The only exception is in the HIF-2α PAS-B G-strand, which shows a high degree of flexibility in its central residues. This is specific of the HIF-2α PAS-

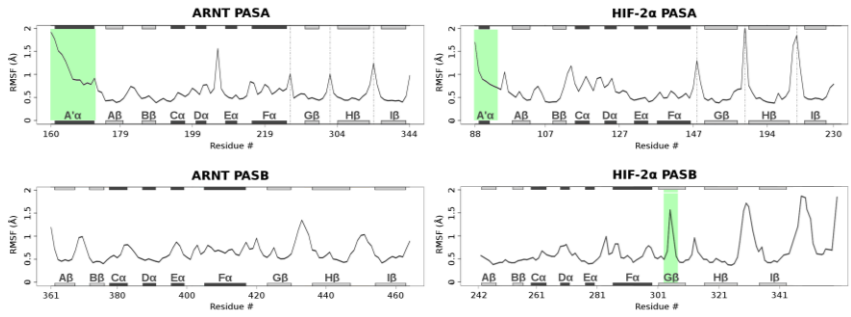

Figure 4.6: RMSF plots for each PAS domain in the dimer. RMSF values are calculated on the Cα atoms. The structured regions with higher fluctuations are highlighted in green and discussed in text. The long and highly flexible PAS-A loops were excluded from calculation and are indicated by dash lines. Helices and β-strands are represented as black and grey bars, respectively, and labelled according to Figure 4.3.

B domain and is not found in the other PAS domains. Moreover, this β-strand is partially unstable and its central region alternates between unstructured (≈ 70%) and folded (≈ 30%) conformations during the simulations (Figure 4.7). This instability of the Gβ strand is consistent with the structures in the NMR ensemble of the isolated HIF-2α PAS-B (PDB ID: 1P97) which contains a completely structured G-strand only in 7 out of 20 states (Figure App. B5). These results highlight an internal rigidity in the PAS domains, with the exception of the highly flexible loops in the PAS-A domains. This suggests that the dynamic of the system is mainly involving quaternary structure oscillations affecting inter-domain interfaces. To

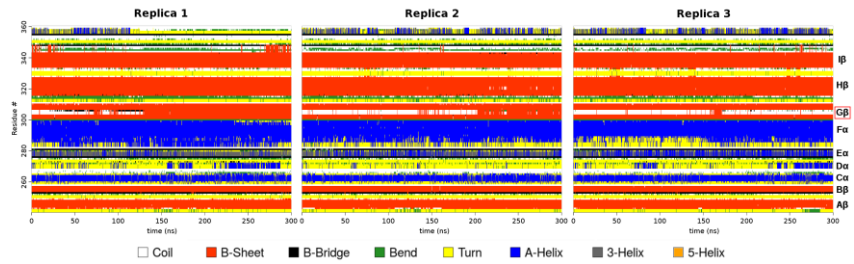

Figure 4.7: Time evolution of the HIF-2α PAS-B secondary structures assignment in the MD simulations of the unbound HIF-2α:ARNT structure. Secondary structure elements were assigned using the DSSP algorithm. The location of the Gβ element is indicated on the right-hand side.

investigate this and to energetically characterize the system interfaces at the residue level, we used the Molecular Mechanics – Generalized Born Surface Area[230,231] (MM-GBSA) method implemented in Amber[232]. This method returns an estimation of the binding free energy associated to a protein-ligand or protein-protein complex formation and it can be conveniently decomposed in contributions to the residue level. A summarized view of the relative contribution to the dimerization provided by each domain is shown in Figure 4.8 and the detailed values are reported in Table App. B1. Values were derived as sum of per-residue contributions averaged over the three replicas. The bHLH domains of the two units equally contribute to the stabilization of the dimer. Due to its central position within the quaternary assembly of the dimer, the HIF-2α PAS-B domain highly contributes to the dimerization free energy, while the ARNT PAS-B domain only interacts with the HIF-2α PAS-B domain, and seems less important in the dimerization. Interestingly, all known inhibitors of dimerization bind to the HIF-2α PAS-B internal cavity,[82–84,212,213] so it is conceivable that perturbing its dynamics could seriously affect the system stability.

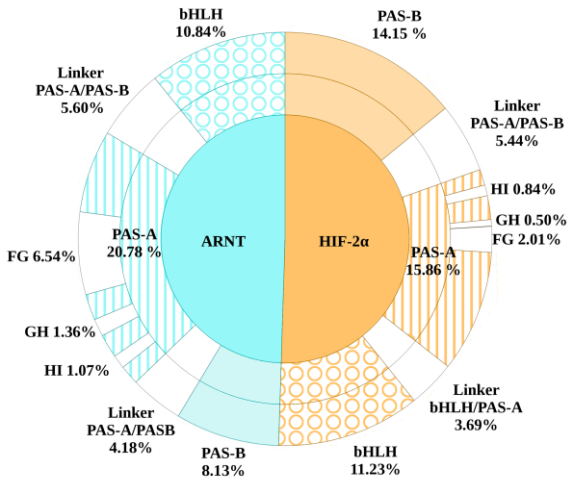

Figure 4.8: Pie-chart illustrating contribution of domains and connective loop/linkers to the total dimerization free energy. ARNT coloured in cyan, HIF-2α in orange.

A feature of PAS-B:PAS-B interface is the involvement of the HIF-2α C-terminal linker, that highly contributes to the dimerization free energy. As expected, the dominant role in the dimer association is adopted by the ARNT PAS-A domain, that interacts with the bHLH region and with both the HIF-2α PAS domains. A relevant insight arising from the MM-GBSA analysis concerns the importance of ARNT PAS-A FG loop. The 30-residue long loop stands out from the other PAS-A loops for its contribution to dimerization which is at least four times greater than the others. This is due to strong interactions with both the C terminal portion of the HIF-2α PAS-A-PAS-B linker and two HIF-2α PAS-B elements, A-strand and C-helix (both highly conserved). This information, associated to the highly-conserved sequence of the ARNT PAS-A FG loop, can suggest a functional role for this loop which enhances ARNT efficacy in binding its partners.

Analysis of the effect of 0X3 inhibitor on the dynamics of the dimer

The dynamics of the HIF-2α:ARNT dimer in the 0X3-bound form (PDB ID 4ZQD) is here presented and compared to that of the apo system (see previous subsection) to identify the regions perturbed by the ligand and shed light on the inhibition mechanism. Three replicas of the system were simulated for 300 ns. Similar to the apo state, the 0X3-bound form has limited flexibility, mostly located in the loop regions.

To investigate inhibitor-induced perturbations of the intra-domain flexibility, the RMSF (after fitting of the Cα atoms of the single domains independently) was calculated (Figure 4.9).

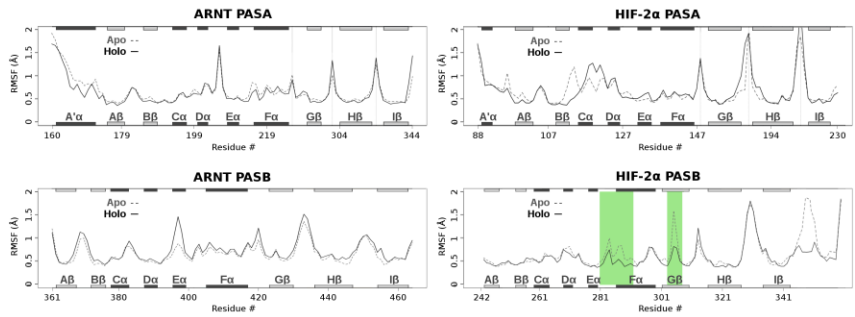

Figure 4.9: Comparison of the RMSF plots for each PAS domain of the HIF-2α:ARNT system in the apo (dashed line) and holo (solid line) simulations. Ligand-perturbed regions discussed in the text are highlighted in light green. The long and highly flexible PAS-A loops are excluded from the calculation, and the corresponding gaps are indicated in the figure by vertical dashed lines. Secondary structure elements according to DSSP are reported on the top and bottom of the graphs (black: α-helix, light grey: β-strand).

The holo and apo simulations show a similar intra-domain RMSF profile for all the domains with some differences in the Cα-Dα-Eα helices region of ARNT PAS-B and HIF-2α PAS-A. These regions are quite flexible and do not contribute to protein-protein interactions. On the other hand, significant differences appear on HIF-2α PAS-B domain in the F-helix and G-strand elements.

These regions are in strong contact with the ligand, and thus probably subjected to a local perturbation. Even the HIF-2α PAS-B C-terminal linker is subjected to a strong rigidification, probably correlated with that of the interacting G-strand.

A comparative analysis of the dimerization free energy at the interfaces was done using the MM-GBSA method. The calculation was performed on the common ensemble of conformations in the region of the principal component (PC) subspace sampled by all three simulation replicas. The PC subspace was calculated for the inter-domain motions (see Par. 4.2 Methods). The per-residue decomposition of the dimerization free energy highlights a weakening of the interactions at the PAS-B:PAS-B interface in presence of the ligand. The most perturbed region involves key residues as ARNT

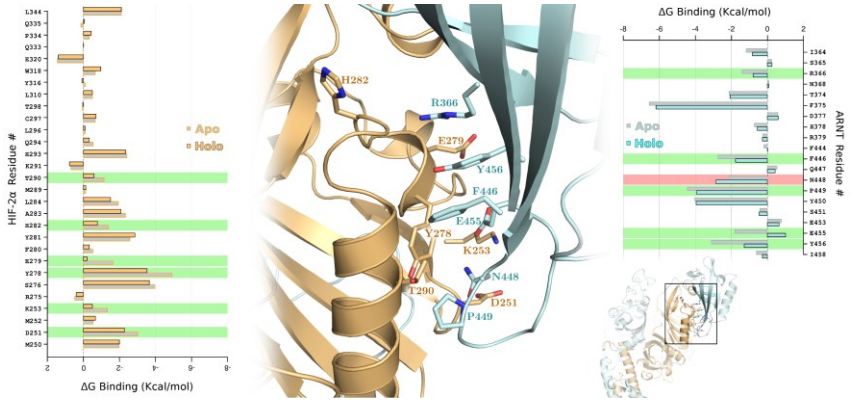

Figure 4.10: Barplot representing the per-residue decomposition of the dimerization free energy calculated with MM-GBSA at the PAS-B:PAS-B interface. On the left HIF-2α residues represented as grey (apo) and orange (holo) bars. On the right ARNT residues represented as grey (apo) and cyan (holo) bars. Residues with apo/holo differences greater than 0.5 (or smaller than -0.5) kcal mol-1 are highlighted with green (or red) background. In the middle panel a 3D representation of the PAS-B:PAS-B interface, with highlighted residues shown in sticks.

R366 and Y456 (Figure 4.10), which are known to be hotspots for dimerization from previous experimental mutagenesis data[85].

The perturbed region of HIF-2α PAS-B includes residues lying on the E and F helices (residues 277-299). As discussed before (Figure 4.9), the arrangement of the F-helix appears to be affected by the presence of the ligand, so we speculate that this perturbation propagates through the helical bundle, destabilizing the whole PAS-B:PAS-B interface. This finding supports the previous hypothesized mechanism proposed by Wu and co-authors based on small perturbation observed in the X-ray structure of the 0X3-bound with respect to the unbound dimer at this interface (PAS-B: Cα RMSD = 0.615 Å, heavy atoms RMSD = 1.207 Å).[85] All the other inter-domain interfaces have association free energies similar to the apo form.

The central part of the HIF-2α PAS-B G-strand is highly flexible and partially unstructured in the apo simulation (see Figure 4.7), while in presence of the 0X3 ligand it is more rigid and fully structured in a β-strand during the entire simulation (Figure 4.11).

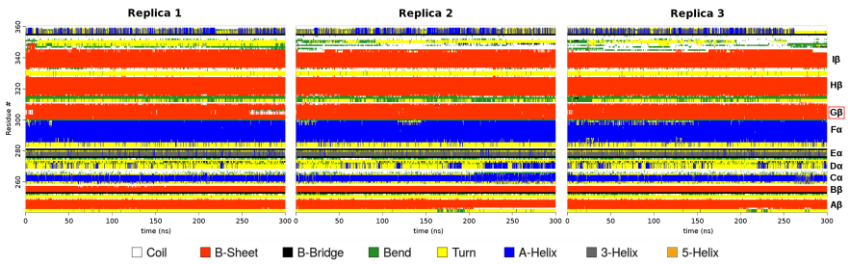

Figure 4.11: Time evolution of HIF-2α PAS-B secondary structure assignment in the MD simulation of the 0X3-bound HIF-2α:ARNT structure. Secondary structure elements were assigned using the DSSP algorithm. The location of the Gβ element is indicated on the right-hand side.

This region of the G-strand includes residues S304, G305 and Q306. S304 is a highly-conserved residue whose side-chain lies within the HIF-2α PAS-B cavity in contact with the ligand. Its importance is also attested by previous mutagenesis experiments reporting that the S304M mutant is unable to bind 0X3 and other similar ligands[83].

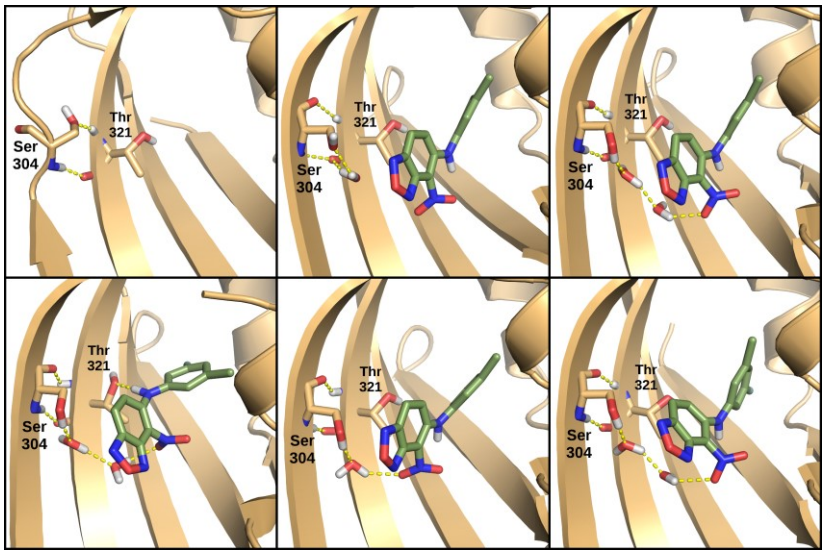

Figure 4.12: Water mediated hydrogen-bond network between Ser304 and the 0X3 ligand. In the unbound form (top left panel) the Ser304 sidechain interacts with the Thr321 backbone, maintaining this region of the G-strand unstructured. In most of the representative structures of the inhibitor-bound form (remaining panels) the Ser304 sidechain is involved in a water-mediated hydrogen-bond with the ligand, and the H-bonds between the Ser304 and Thr321 backbones facilitate a complete structuring of the strand.

Visual inspection of this region during the simulations highlighted a stable hydrogen-bond network between S304 and the nitro group of the ligand mediated by one or two water molecules. A set of representative arrangements of these water-bridged interactions (Figure 4.12) was extracted by cluster analysis (see Par. 4.2 Methods).

The presence of these interactions looks essential for the complete folding of the G-strand. Moreover, residue Q306 is known to interact with the proflavin inhibitor (Figure 4.13) that, in a reported crystal structure (PDB ID: 4ZPH),[85] is shown to bind outside the PAS-B ligand binding cavity known for the bHLH-PAS proteins.[209] This local perturbation does not completely justify the expected lower stability of the dimer in presence of 0X3. To investigate the long-distance effects of this local perturbation we analysed residue correlated motions by means of the distance cross correlation matrix[64,236] (DCCM) calculated on the Cα carbons of the concatenated trajectories (Figure 4.14). Each domain of the system holds strong

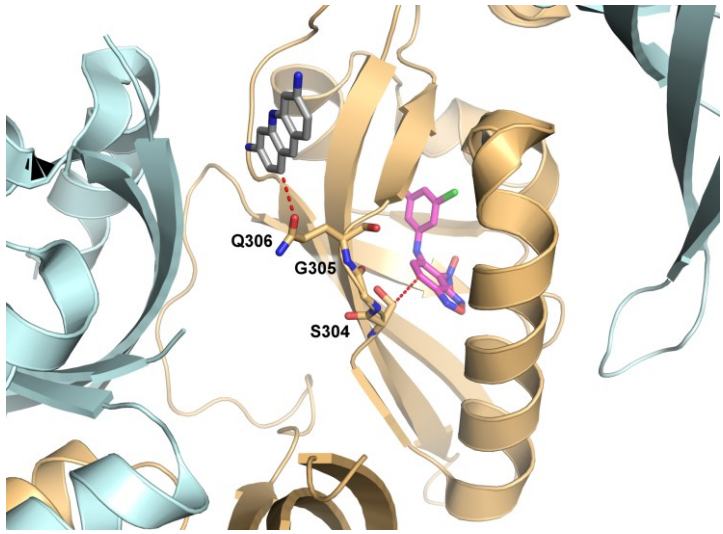

Figure 4.13: Close-up of the HIF-2α PAS-B region around the 0X3 ligand from the X-ray structure. In the cartoon representation, HIF-2α PAS-B is in orange, ARNT PAS-A in cyan. HIF-2α residues in the PAS-B G-strand that are involved in the ligand-protein interactions are shown as sticks. In addition to the 0X3 inhibitor from the 4ZQD structure (in purple sticks), also the position of proflavin from the 4ZPH structure (in grey sticks) is represented.

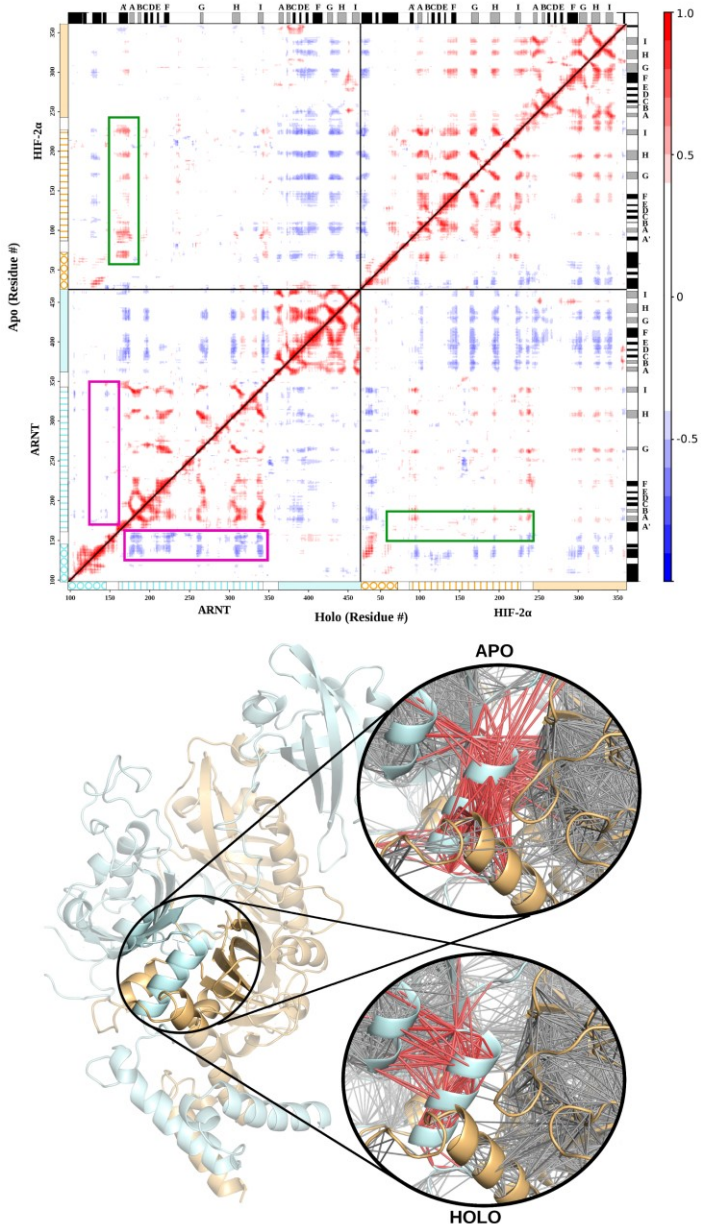

Figure 4.14: Distance cross correlation matrices for apo and holo simulations. The correlation matrices are shown on the top (upper triangular for apo – lower for holo). Domains (circle=bHLH, vertical lines=PAS-A, light filled=PAS-B) are labelled on the left and on the bottom. Secondary structure profiles (black=helix, light-grey=sheet) are labelled on the top and on the right. A 3D representation of the network derived from the DCCM is represented at the bottom with a close-up showing the differences between the apo and holo connections in the ARNT PAS-A A' helix region.

internal positive correlation (except for the long PAS-A loops) in both apo and holo simulations, confirming the rigidity of all the PAS domains during the simulations. Major differences are evident in the region of ARNT residues 130-180. In particular, ARNT bHLH-PAS-A linker (magenta squares in Figure 4.14) has anti-correlated motions towards the ARNT PAS-A domain only in the holo simulation, while in the apo simulation the ARNT PAS-A A' helix (green squares in Figure 4.14) is correlated with the HIF-2α PAS-A domain. This suggests that the motion of ARNT PAS-A A' helix, lying at the PAS-A:PAS-A interface, becomes decoupled from the PAS-A domain after inhibitor binding, probably indicating lower inter-domain interaction. To correlate the altered flexibility of HIF-2α PAS-B G-strand with the perturbed dynamics and correlated motions of ARNT PAS-A A' helix, we calculated the optimal and suboptimal paths[245] between these two regions from the DCCM networks of the apo and holo simulations (Figure 4.15). In the case of the apo network, the shortest paths connect the HIF-2α PAS-B G-strand with the HIF-2α PAS-A strands and then with the ARNT PAS-A A' helix (Figure 4.16). In the case of the holo network, the shortest path is altered and links

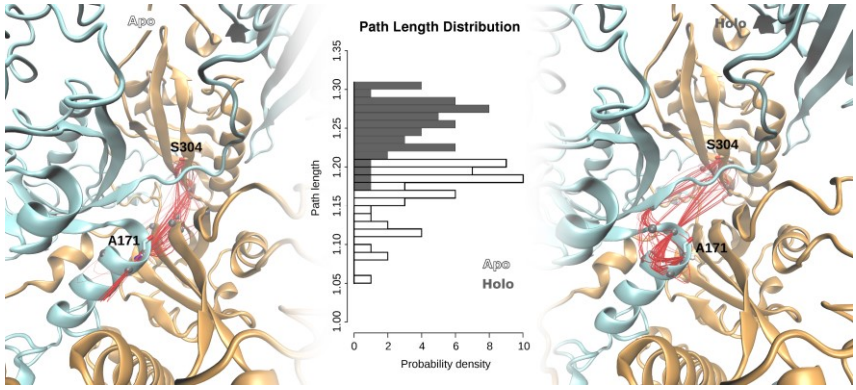

Figure 4.15: Optimal and suboptimal path analysis for the unbound (apo) and 0X3-bound (holo) HIF-2α:ARNT dimer structures. Left and right panels: paths are shown as red lines connecting residues in the three interface domains (HIF-2α PAS-B and PAS-A in orange cartoon, ARNT PAS-A in cyan cartoon). Central panel: path length distributions between the HIF-2α S304 and ARNT A171 residues. Apo: empty bars; holo: grey filled bars.

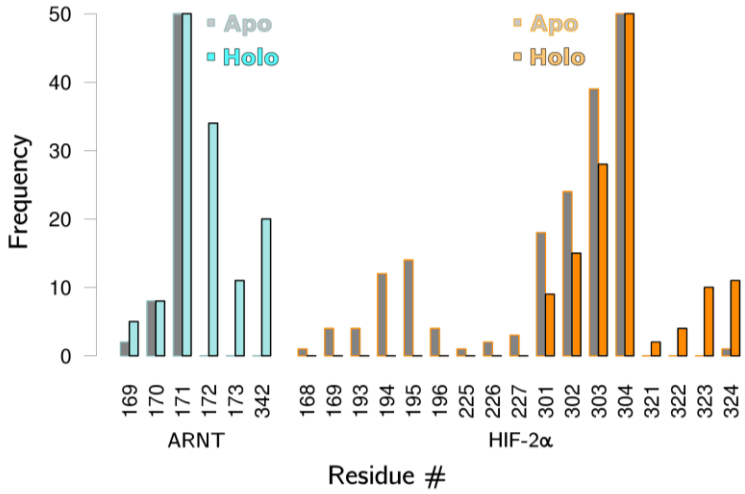

Figure 4.16: Comparison of residues in suboptimal communication paths in the apo and holo simulations. The frequency of each residue occurrence in the best 50 suboptimal paths is shown.

the ligand-perturbed HIF-2α PAS-B G-strand to the ARNT PAS-A, implying a longer connection to the A' helix. These results suggest that perturbation of the HIF-2α PAS-B G-strand by the ligand has an impact on the PAS domain-domain interaction and could affect the dimer stability.

4.4 Discussion

Since the discovery of a large cavity within the PAS-B domain of HIF-2α and the identification of compounds that bind this cavity and dissociate HIF-2α from ARNT,[82,84,213] several structure-based research programs have been started to find selective and potent antagonists of the HIF-2α transcriptional activity.[83,212,246,247] However, mechanistic understanding of ligand effects on the dimer association had remained elusive until recently, because the available X-ray structures of HIF-2α in complex with artificial ligands encompassed only the isolated HIF-2α and ARNT PAS-B domains. It was first suggested that ligands can induce conformational changes

at the PAS-B β-sheet of HIF-2α that weaken the interactions with the ARNT PAS-B β-sheet,[83,213] but no evidence in the context of the full dimer was available. Only recently the determination of the crystallographic structures of the entire bHLH-PAS region of the dimer in the apo and 0X3-bound forms[85] has opened the way to a better understanding of the inhibition mechanism. The proximity of 0X3 to the α-helices region of the HIF-2α PAS-B domain supports a model in which ligand binding could influence the heterodimer stability through a perturbation of its PAS-B:PAS-B interface.[85] However, deeper insight in the atomistic details of this perturbation is limited by the static view provided by the crystallographic structure. Indeed, it is conceivable that 0X3 perturbation could propagate through the structure and affect other interfaces thanks to the dimer intrinsic dynamics and in agreement with a previously suggested allosteric inhibition mechanism.[83,85] To investigate this hypothesis, we characterised the evolutionary, dynamical and energetic properties of the dimerization interfaces in the apo and 0X3-bound form. The results shed light on the atomistic details of 0X3 inhibition mechanism, on the residues involved in dimer stabilisation and on pharmacophoric features required for future development of analogues of 0X3.

Our residue conservation analysis (Figure 4.5) detected high scoring patches on all inter-domain interfaces confirming homomeric and heteromeric interactions in agreement with the crystallographic structure.[85] In addition, our results highlighted strong conservation in two connecting elements that may be involved in ARNT flexible arrangement around different partners (ARNT-PAS-A FG loop) and in the stabilisation of HIF-2α dimerization (HIF-2α-PAS-A/PAS-B linker). Past mutagenesis and co-immunoprecipitation (co-IP) studies indicated that the bHLH:bHLH, PAS-A:PAS-A and HIF-2α PAS-B:ARNT PAS-A interfaces are critical for dimer stability.[85] We assessed this by calculation of the contributions provided by each domain and secondary structure element to the dimerization free-

energy. We confirmed the importance of the bHLH and PAS-A domains of both partners and of the HIF-2α PAS-B domain in the dimer stabilization (Figure 4.8). In addition, we showed that the ARNT-PAS-A domain gives the major contribution to binding and that the dynamic behaviour of the PAS-A FG loop enhances the domain intermolecular interactions by wrapping around the HIF-2α PAS-B domain. Interestingly, for the C-terminal linker of HIF-2α, that is inserted into the PAS-B:PAS-B interface and is partially folded in a α-helix in all the available crystal structures of ARNT dimers (with HIF-1α, NPAS1 and NPAS3),[85,193,205] our analysis evidenced a relevant contribution to protein dimerization.

From MD simulations we detected a general internal rigidity in the PAS domains of both partners, except for the PAS-A loops. This suggests that the dynamics of the system mainly involves quaternary structure oscillations. These motions are particularly evident for the ARNT PAS-B domain, that gives few interactions with the rest of the dimer, and shows characteristic hinge-bending motions around the flexible PAS-A/PAS-B linker. The dynamics of this domain, along with its arrangement (Figure App. B4) in the crystallographic structures of ARNT in complex with different bHLH-PAS class-I partners (HIF-1α, HIF-2α, NPAS1 and NPAS3),[85,191,193,205] supports a general model for ARNT dimerization in different heterodimers: strong interactions at the dimerization interfaces in the bHLH/PAS-A region stabilize the dimerization, while the domain bending motion of ARNT PAS-B provides adaptation to different partners through different dimerization geometries in the PAS-B:PAS-B region.

We compared the dimer dynamics in the apo and 0X3-bound forms and identified the dimerization interfaces that are mainly affected by ligand binding as well as the ligand-induced perturbations on intra-domain correlated motions and on inter-domain communication paths. The holo dimer has reduced flexibility in the Eα-Fα region of the HIF-2α PAS-B domain and weakened residue interactions in the

PAS-B:PAS-B interface (Figure 4.10), thus confirming the hypothesis of Wu and co-workers about the involvement of this interface in dimer inhibition.[85] However, in the holo simulations we also detected a previously undescribed perturbation on the opposite side of the HIF-2α PAS-B domain: the G-strand, which is flexible and partially unstructured in the apo simulation, becomes more rigid and fully structured in a β-strand in the presence of the ligand. This regularisation of the strand is triggered by water-bridged interactions of the 0X3 nitro-group with S304 sidechain (Figure 4.12). Previous studies on the isolated HIF-2α PAS-B domain have demonstrated that, among a number of artificial ligands, the ones with a nitrobenzoxadiazole group connected to aromatic/heterocyclic rings by a amine linker, like 0X3, show the highest binding affinities and inhibition potency.[83] While the heterocycle and the nitro group in this molecular moiety were suggested to contribute to high affinity through favourable electrostatic interactions with a few side-chains in the PAS-B binding cavity,[83] no clear explanation was provided for their role in the inhibition mechanism. On the other hand, a critical role of water molecules in the stabilisation of the apo cavity was previously reported,[248] but our insight on the dynamics of the bound form explains for the first time the atomistic details of 0X3 perturbation mediated by water. Here we propose that the local effect of the inhibitor propagates through HIF-2α-PAS-B interfaces with other domains toward the core dimerization region. This is evident in the change of correlated atomic motions between HIF-2α and ARNT domains. The DCCM network analysis showed that a communication path connecting HIF-2α-PAS-B - HIF-2α-PAS-A - ARNT-PAS-A is present in the apo but lost in the holo simulations (Figure 4.14). The motion decoupling induced by the loss of this communication is consistent with a weakening of the HIF-2α-PAS-A:ARNT-PAS-A interaction in the A' helix key region. A set of residues (in the region 168-227) of the HIF-2α-PAS-A domain (Figure 4.16) are only present in the shortest path of the apo simulations and are expected to be

critical to sustain the dimer interaction. Indeed three of them have been already shown to be essential by previous studies.[85]

These results supported a model of inhibition by 0X3 where both HIF-2α-PAS-B interfaces are destabilised: the helices side interacting with the ARNT-PAS-B β-sheet, and the β-sheet interacting with both the PAS-A domains. This latter perturbation allosterically propagates to the PAS-A:PAS-A interaction interface thus destabilizing one of the most important region for dimer association. A critical role in the initial induction of these effects is played by water-bridged ligand-protein interactions. This suggests that, in addition to previously identified features of successful inhibition of 0X3,[83] future drug design may be targeted to insert functional groups to stabilise water-bridged interactions with the key residues in the HIF-2α-PAS-B G-strand.

"It will be found that everything depends on the composition of the forces with which the particles of matter act upon one another; and from these forces, as a matter of fact, all phenomena of Nature take their origin."

Ruggero Giuseppe (Roger Joseph) Boscovich, - Philosophiae Naturalis Theoria (1758)

CHAPTER 5

INVESTIGATION OF ADENOSINE A2A RECEPTOR DIMERIZATION THROUGH COARSE-GRAINED METADYNAMICS

5.1 Introduction

G-protein coupled receptors (GPCRs) are the largest family of human membrane proteins, and mediate cellular responses to hormones, neurotransmitters, chemokines and the senses of sight, olfaction and taste. Given their importance as sensors of cells, they represent primary targets of about one third of currently marketed

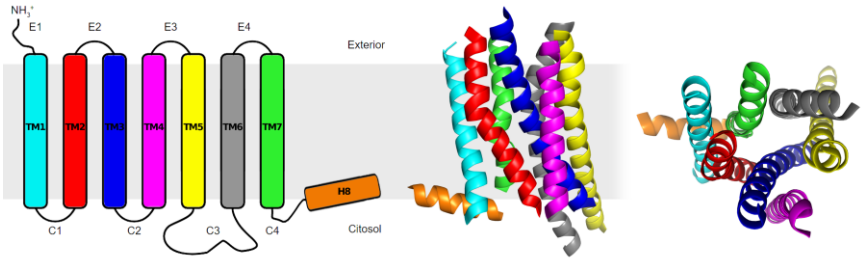

Figure 5.1: General structure of GPCR. The topology diagram (left) and 3D structure from side (center) and extracellular (right) view are represented with different colours for the 8 helices.

drugs for treating many human diseases, including cancer, diabetes, obesity, heart failure and neurological diseases.[249] Despite the diverse array of signals recognized by GPCRs, they all share a common canonical fold of seven transmembrane helical segments (TM1-TM7) connected by three extracellular loops and three intracellular loops, plus an intracellular amphipathic helix (H8) that is not solved in all the GPCRs X-ray structures suggesting a higher conformational flexibility in that region for some GPCRs (Figure 5.1).[250]

All GPCRs reside in the cell membrane, which has a remarkably complicated and heterogeneous architecture: it consists of a variety of glycerophospholipids, sphingolipids, cholesterol and membrane proteins. The general activation mechanism of GPCR starts with a conformational change, mainly involving the intracellular side of TM5 and TM6, upon external signal (e.g. ligands) that allows *G proteins*, *arrestins* and other signalling proteins to bind to a GPCR's intracellular surface.[251] This control mechanism is highly complex; different ligands can indeed stimulate different intracellular signalling pathways independently through a single GPCR, and many GPCRs possess multiple ligand-binding sites that influence intracellular signalling in distinct manners.[252] In Figure 5.2 is depicted a general G protein activation mechanism that elicits the production of second messengers. In the first stage the signalling molecule (e.g., hormones, neurotransmitters) binds to the GPCR from the extracellular side of

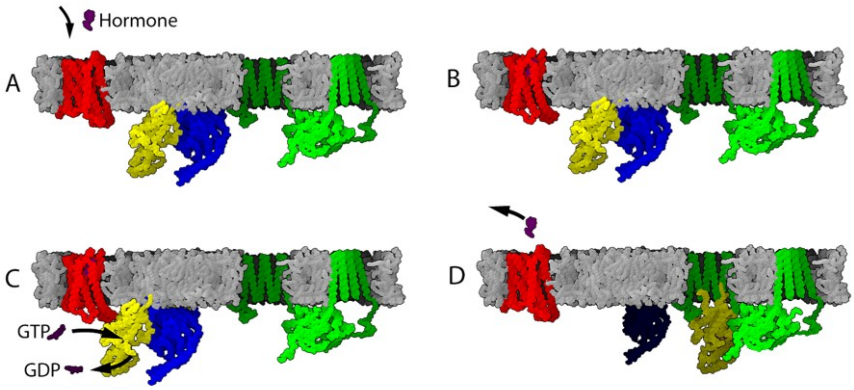

Figure 5.2: GPCR activation mechanism by a generic hormone. A) The hormone molecule binds to the GPCR (red) from the extracellular side of the membrane. B) Binding stimulates a GPCR conformational change in the intracellular side. C) G protein α (yellow) and βγ (blue) subunits bind to GPCR, promoting the GDP/GTP exchange within the α subunit. D) The Gα subunit is released and activate the specific signal cascade. (Image taken from Wikipedia)

the membrane, stimulating a conformational change in the intracellular side of the receptor that promotes the exchange of a molecule of GDP for GTP at the α subunit of the G protein. The latter represents the onset of the signal cascade (Figure 2). The final effect depends on the type of G-protein bound to the receptor (i.e., $G_{\alpha s}/G_{\alpha i/o}$, $G_{\alpha q/11}$ or $G_{\alpha 12/13}$).

In addition to cell signalling modulated by the G_α subunit, the βγ subunits can also regulate different signalling pathways. Therefore, understanding the structural basis of the G protein functional mechanism is of pivotal importance for the development of new drugs able to modulate specific GPCRs and have the desired therapeutic effect. Although GPCRs were initially thought to function exclusively as monomeric entities, evidences in the last decades indicate that they can form homomers and heteromers in intact cells.[253–256] These data come from various biophysical techniques, such bioluminescence resonance energy transfer (BRET), fluorescence resonance energy transfer (FRET), fluorescence complementation or combination of these techniques. Even multi assembly (dimers or tetramers) X-ray

structures of GPCR are available and have been used to characterize protein-protein interfaces.[250,257–260] However, given the particular conditions employed during crystallization, it is possible that some of the observed interfaces may be affected by experimental condition such as crystal packing effects.[256]

Nevertheless, some interfaces have been observed more often than others: TM5 and TM6 residues constitute the main interfaces for chemokine CXCR4[250] and μ-opioid[258] receptor crystallized dimers, while involvement of TM6 was also suggested for β2-adrenoceptor dimers[261] and leukotriene receptor[262], and TM5 for homodimerization of dopamine D2[263], muscarinic M3[264] and serotonin 5-HT_{2C}.[265] Apart from the TM5-TM6 interface, crystallized chemokine CXCR4 dimers also show contacts at the intracellular ends of TM3 and TM4,[250] while studies on μ-opioid dimers indicate a second, less prominent symmetric interface, involving TM1, TM2, and H8[258]. A TM1–TM2–H8 interface was also found in crystals of k-opioid receptor dimers[257], rhodopsin[266,267], opsin[268] and β1-adrenoceptor[259] (the last with an additional interface involving TM4 and TM5). Notably, the two crystallographic interfaces of the β1-adrenoceptor were suggested to be physiologically relevant with cysteine-cross-linking experiments.[259]

It has also been suggested that the association of different GPCRs can influence the activation process.[258,259] The preferred association interfaces can also vary depending on the ligand-induced conformation of the receptor, therefore it is important to assess if different ligands can also promote different dimeric interfaces by stabilizing different receptor conformations. Moreover, in different studies a negative cooperativity between the protomers of a dimer has been observed.[269–271] Cooperativity (positive or negative) is a particular type of allosteric modulation in receptor oligomers, in which the protomers are the conduit of the allosteric modulation and the same ligand can work both as allosteric modulator (binding to the first protomer) and the target modulator (binding to the second

protomer). It has been proposed that negative cooperativity could provide a mechanism that protects the biologic system against over activation of endogenous ligand.[272]

Within the intricate landscape of cooperativity, understanding the association process of GPCRs at atomistic level becomes of pivotal importance. MD simulations and related molecular simulation approaches provide important tools which allow to simulate both individual membrane proteins and more complex membrane systems. Given the complexity of the dimerization process, which can require from microseconds to seconds, a coarse grained (CG) approach is often used to simplify the representation of the system. The approaches used to study the association process can be classified in two groups: *self-assembly simulations* and *biased simulations*.[273] Self-assembly simulations consist in embedding a large number of receptors in a preformed and regularly spaced lipid bilayer to maximize their dispersion in the membrane; this is often performed repeating a unit cell that contains a single receptor. From that dispersed configuration it is possible to follow the assembly of the receptors. Unfortunately, in the time-scale accessible to the CG MD simulations, the event typically observed is the association of the receptors, with only limited events of dissociation. Therefore, using this kind of simulations one can obtain only partial description on the receptor behaviour in a lipid bilayer, such as their propensity to oligomerize, the accessible interfaces, and the dynamics of the contacts. This kind of simulations, performed with different conditions, has been used to characterize several oligomerization processes.[56,57,274–279] In principle, the relative strength of protein interfaces can be extracted from the Potential of Mean Force (PMF) computed from a conventional unbiased simulation. In order to achieve convergence of the calculation, during the simulation the system should sample a full exploration of the conformational space at equilibrium, and in the case of binding and unbinding events the convergence is not yet accessible in the current time-scale of CG

simulations. A viable alternative is to determine the PMF between two receptors using biased simulation approaches such as Umbrella Sampling (US) or metadynamics. In particular, US has been widely used to compute the free energy of association for specific interfaces. However, for the sake of the convergence of the calculation, using US the user generally chooses the most relevant degree of freedom of the system, typically the distance between protomers, neglecting other relevant properties such as orientation of the protomers one relative to the other.[278,280,281] To deal with the high dimensionality of the problem, it has been recently proposed to combine US and metadynamics, which provide a more accurate exploration of the free energy landscape. In this approach, in addition to the constant external harmonic bias of the US algorithm applied to the distance collective variable (CV), a time-dependent sum of Gaussian bias (metadynamics) was applied to enhance the sampling of the angle that defines the orientation of one promoter relative to the other.[282]

In the present work the association process of Adenosine A2A receptor (A2aR) is investigated using a recently reported innovative protocol called Coarse-Grained MetaDynamics (CG-MetaD),[283] which combines CG molecular dynamics and well-tempered metadynamics (see Par. 2.7 Metadynamics). A2aR is a GPCR ubiquitously expressed in the body that plays a pivotal role both in important normal response functions (*e.g.* inflammation) and under pathological conditions (e.g. Parkinson's disease and attention deficit hyperactivity disorder). A2aR is capable of responding to an impressive plethora of stimuli (*e.g.* light, hormones, neurotransmitters) and its activation results in the elevation of intracellular cAMP.[284] However, the actions of A2aR are complicated by the fact that a variety of functional heteromers, composed of a mixture of A2aR subunits with subunits from other unrelated GPCRs, have been found in the brain, thus adding a further degree of complexity to the role of adenosine in modulation of neuronal activity. Heteromers with adenosine A1,[285] dopamine D2[286] and D3,[287]

glutamate mGluR5[288] and cannabinoid CB1[289] have been observed, as well as CB1/A2aR/D2 heterotrimers,[290] but the functional significance and the endogenous role of these hybrid receptors are still to be unravelled. Moreover, A2aR was proven to form homodimers within the cell membrane using FRET and BRET experiments, and it was demonstrated that homodimers but not single protomers are the functional species at the cell surface.[291] To gain insight into the homomeric association of A2aR, we performed extensive CG-MetaD simulations using the Martini[109] CG force field and the Multiple Walker (MW) protocol[126], the latter used to further enhance the exploration of the free energy landscape. The CG-MetaD approach has been recently employed to describe the association process of the transmembrane domain of the epidermal growth factor receptor.[283] Overall, we have performed 1 ms of CG-MetaD simulation in which we have observed several association and dissociation events between the A2aR protomers, thus providing an accurate and comprehensive description of the associated free-energy surface (FES). This allowed identification of the lowest free-energy basins that correspond to distinctive conformations of A2aR homodimer.

5.2 Methods

Coarse-grained model

The starting conformation of the A2aR has been taken from the 2YDO X-ray structure.[292] The structure has missing residues that were modelled with the MODELLER[293] software using the 3RFM structure[294] as template. The atomistic structure was first converted to the MARTINI force-field using the *martinize* tool,[111] using the ELNEDYN representation (see Par. 2.5 Coarse-Grained Molecular Dynamics) to mimic the α-helical hydrogen bonds. In the next step

two A2aR protomers were placed by *insane*[295] at the distance of about 7 nm and inserted in a squared CG lipid bilayer with side of 20 nm composed of *1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine* (POPC)/cholesterol with 9:1 ratio. This system was then subjected to 10000 steps of steepest descent minimization, followed by four runs of 10000 steps each, at increasing timesteps: 1 fs, 2 fs, 4 fs 10 fs and a final run of 20 ns at 20 fs time-step. All these steps were performed in NVT ensemble coupling the temperature at 300 K using the v-rescale thermostat with a 1ps time constant, while the protein was restrained to the starting coordinate with a 1000 kJ mol^{-1} nm^{-2} force constant. The electrostatic interactions were shifted to 0 between 0 and 1.1 nm. The van der Waals forces were described by the 12-6 Lennard-Jones potential that was shifted to zero between 0.9 and 1.1 nm. The ensemble was then switched to NPT and pressure was controlled in a semi-isotropical manner (xy and z were independent) to 1 bar using the Berendsen barostat with a 5 ps time-constant. The system was simulated in these conditions for three runs of 5 ns each, decreasing the force constant from 1000, to 500 and 0 kJ mol^{-1} nm^{-2}. All simulations were performed using Gromacs 5.1.

Coarse-grained metadynamics

The CG-MetaD simulation was performed at 300 K with a time step of 20 fs, using the WT version of metadynamics (see Par. 2.7 Metadynamics), implemented in Plumed 2.3. Ten parallel simulations were performed according to the multiple walkers (MW) approach (see Par. 2.7 Metadynamics), each started from the equilibrated conformation of the system. The distance between the two proteins (r) and a torsion that describes the reciprocal orientation (Ω) were chosen as active collective variables. The distance r is defined as the distance between the center of mass of backbone beads of residues A20, A51, D52, Q89, S90, S91, I92, F93, C128, N181, F242, A243,

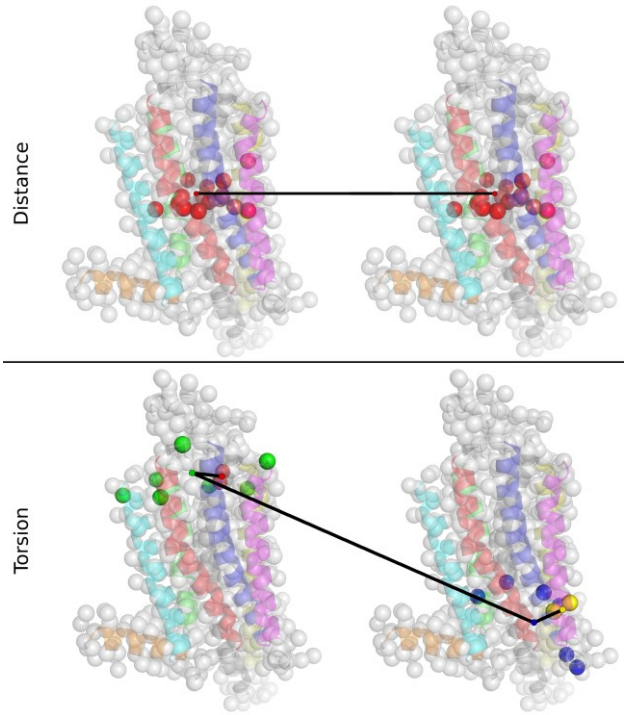

Figure 5.3: Geometrical definition of CVs represented on the CG protomers (grey spheres) in which the secondary structure elements are reported with coloured cartoons. Upper panel: distance r between the center of mass of selected backbone beads; Lower panel: torsion Ω between four center of mass points (represented as red, green, blue and yellow spheres) that define the protomer reciprocal orientation.

C245, N280, S281 on each protomer (Figure 5.3) and represents the distance between the two proteins. The torsion Ω is defined as the dihedral angle among four points, computed as the center of mass of backbone beads (Figure 5.3). Residues included for calculation of the four points are: N253, F255 for the first point; V8, I64, A72, G142, V178, I252, W268 for the second point; A97, P109, L110, G123, R199, I238, F286 for the third point; K122, G123 for the fourth point. These two CVs allowed us to efficiently explore the FES of the system. Gaussians of height 0.5 kJ/mol and width 0.04 nm for r and 0.06 rad for Ω were used and deposited every 5000 steps with a bias factor of 20 for each walker. An upper wall limit was imposed for the r CV at a value of 8 nm to limit the exploration of unbound states.

Analysis

The free energy landscape of the system was obtained from the metadynamics calculation using the plumed *sum hills* tool. The simulations of the 10 walkers were concatenated with a time-step of 10 ns for a total of 100000 frames, and conformations lying within a region that encompass each minimum were extracted. The regions used for the three minima definition are reported in Table 5.1.

Table 5.1: Geometric definition of bound free-energy minima

Minimum	*r* min (nm)	*r* max (nm)	Ω min (°)	Ω max (°)
M1	3.16	3.49	-82	-18
M2	3.56	3.94	-70	-5
M3	2.97	3.35	170	-161

Frames belonging to each minimum were clustered with the GROMOS clustering method based on the backbone beads, with a distance cut-off of 2.0 nm. The centrotype of the most populated cluster was selected as representative of the minimum conformation. It was then back-mapped to atomistic resolution to better characterize the dimeric interfaces using the backward script provided by the MARTINI developers[296] with the CHARMM36 force-field.

The density distribution of cholesterol molecules was calculated dividing the space on the xy plane in a 150x150 grid and counting the number of molecules within each bin. The resulting densities were then normalized by the value that would be obtained for a uniform distribution of cholesterols.

MD trajectories were visually inspected using the Visual Molecular dynamics (VMD) software,[177] while structural analysis of representative conformations and image generation were performed with the Chimera software.[297]

5.3 Results

Sampling of the A2aR dimerization took a total of 1 ms of CG-MetaD simulation (100 μs for each walker), during which the evolution of the FES as a function of the two CVs was assessed to evaluate the convergence. After 800 μs the FES remained unchanged, while the system continues exploring the whole phase space. At this condition, we consider the simulation converged (Figure 5.4).

From the FES one can accurately quantify free energies of interaction. Thus, the free energy of dimerization was calculated from the FES as the difference between the dimeric (conformations at 3nm $<$ r $<$ 4nm) and monomeric (conformations at r $>$ 6.5 nm) states, yielding an estimate of -85.2 ± 1.2 kJ/mol (error calculated as the standard deviation from the average value of the TM helix dimerization free energy obtained from the last 250 μs where the calculation is converged) (Figure 5.5). The large energy difference between bound and unbound states computed from our metadynamics simulation seems to overestimate the stability of the dimeric state, which may be ascribed to a bias of MARTINI force-field towards dimeric assembly of proteins. It should be noted that previous result from FRET experiments showed that more than 90% of A2aR are in homo-dimeric state within the cell membrane, suggesting a high free energy of association for the homodimerization.[291]

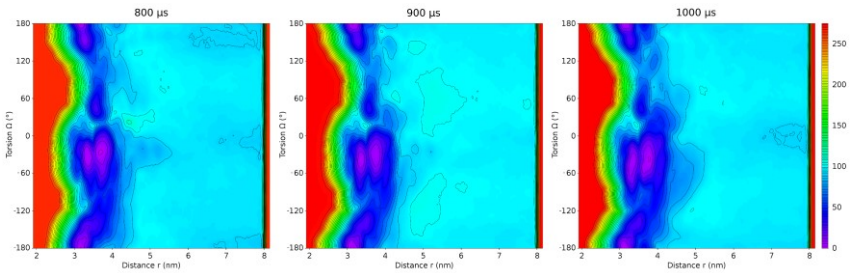

Figure 5.4: The FES of the system for the two selected CVs was evaluated during the evolution of the simulation. Here we report the last 200 μs of simulation during which the shape of FES undergoes negligible changes.

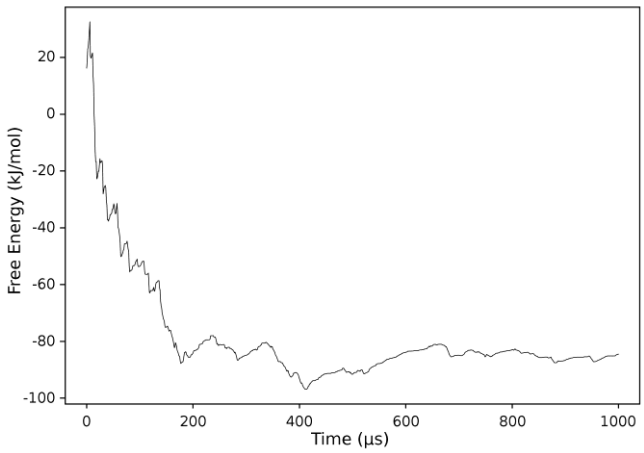

Figure 5.5: Plot of free-energy difference between the bound (3nm < r < 4nm) and unbound (r > 6.5nm) states as a function of time. The free- energy difference was calculated as the difference between the bottom of the well of the dimeric states and the plateau of the monomeric states in the mono-dimensional FES (for the distance CV).

Despite, several forth and back events were observed between the dimeric and monomeric states (Figure 5.6). We note that some walkers explored only part of the phase space (as for walker 7 and 8), while others visited the whole CV space sampling several binding/unbinding events. This effect suggests a not completely converged simulation, and may be due to specific degrees of freedom

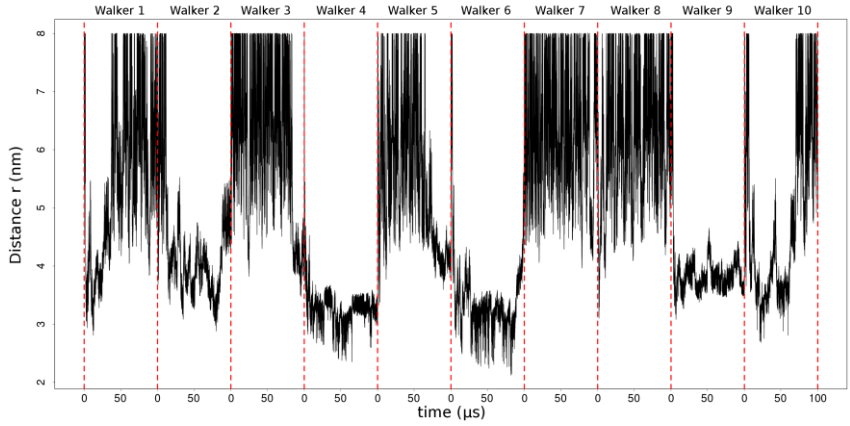

Figure 5.6: Monitoring of the distance CV (r) during the simulations of the 10 walkers.

of the system (e.g., contacts between GPCRs atoms with lipids or cholesterol) that are not taken into account by effective but coarse-grained representation of the phase space using the distance and torsion CV. The use of the multiple-walkers approach allows to increase the total simulation time, parallel growing the bias in different regions of the FES. This caused a rapid exploration of the FES but limits the sampling of binding/unbinding events. Overall, we believe that the large simulation time reached allowed to overcome these limitations.

The final FES (Figure 5.7) shows isoenergetic unbound states at higher energy (about 85 kJ/mol) compared to the lower energy bound state. The first minimum (indicated as M1 in Figure 5.7) lies at a distance r of about 3.3 nm and values of the torsion Ω between -20° and -80°. The conformations belonging to this minimum present symmetric interaction involving the TM1-TM2-H8 interface. The second minimum (indicated as M2 in Figure 5.7) was found at a larger r (about 3.8 nm), but at the same Ω value. The dimerization interfaces of the conformations belonging to this minimum are indeed the same of M1, with the difference that there are weaker contacts between the two protomers in the central part of the interface, resulting in a larger distance value. The third minimum (indicated as M3 in Figure 5.7) was found at r values of about 3.2 nm and at Ω values between 160° and -160°. In this case the dimerization interface is asymmetric and involve TM1-TM2-H8 of one protomer and TM3-TM4-TM5 of the second protomer.

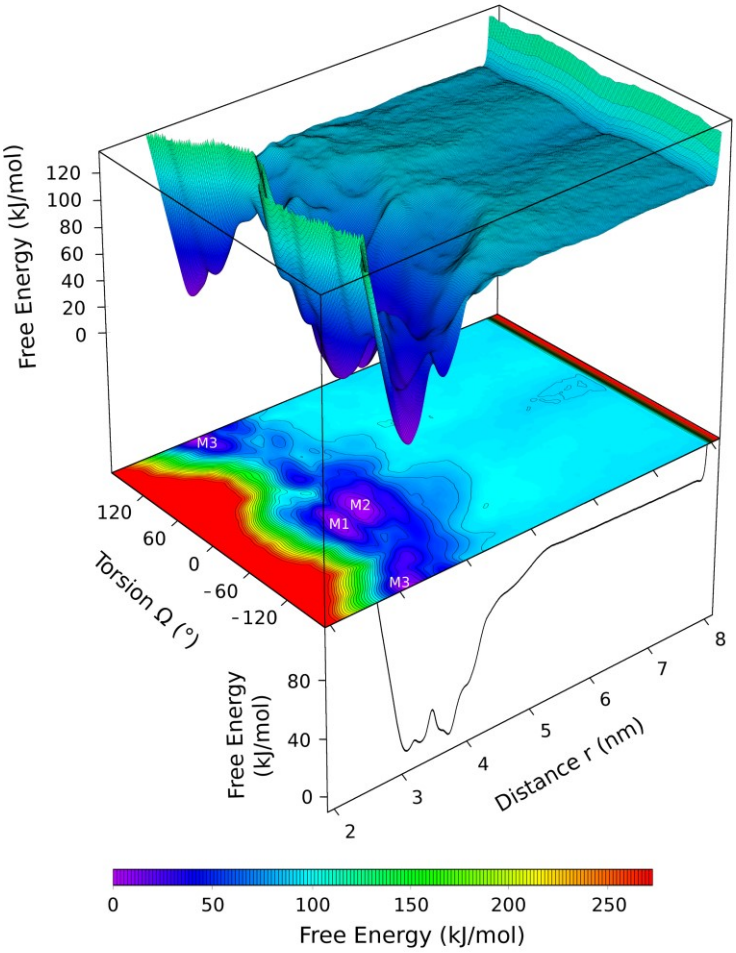

Figure 5.7: Final FES. 3D-view with colours that range from purple to red (upper panel). Projection of the FES in the r and Ω dimensions represented with isoenergetic lines (middle panel). One-dimensional free-energy profile as a function of the distance CV (lower panel).

The GPCR protomers are embedded in the membrane bilayer formed by POPC endowed with 10% of cholesterol molecules. To better characterize the role of cholesterol molecules in stabilizing bound and unbound conformations, we mapped the cholesterol density on the xy plane around the proteins for each state (unbound, M1, M2, M3). The resulting distribution maps (Figure 5.8) show

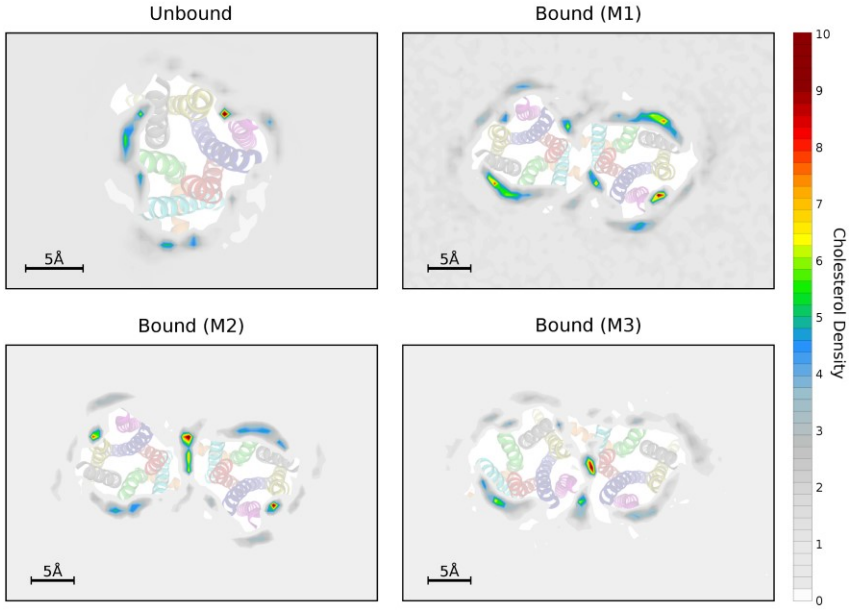

Figure 5.8: Cholesterol density map on the xy plane for the unbound (upper left panel), bound M1 (upper right panel), bound M2 (lower left panel), and bound M3 (lower right panel) states. Low cholesterol densities are reported in shades of grey, (with value of 1 for the uniform distribution) while spots of higher densities are represented with colours ranging from cyan to green, yellow and dark red. The A2aR protomers are represented as cartoon coloured according to Figure 5.1.

regions with high density of cholesterol, indicating several sites of interactions between proteins and cholesterol. In the unbound state the main interaction site for cholesterol was found between TM4 and TM5 helices, while lower densities were found also in correspondence of TM6-TM7 and TM1-H8. Interestingly, in the bound states the protein-protein interfaces create new pockets in addition to the cholesterol binding sites found for the unbound states. In M2 and M3, in particular, cholesterol mediates protein-protein interactions by interposing between the two protomers, while in the M1 state, where the proteins are in tight contact, no additional cholesterol density was found at the interface. It should be noted that the cholesterol density distributions do not significantly change in the different minima, except for the protein-protein interfaces.

To better characterize the dimeric interfaces, the representative CG structures of each minimum were back-mapped to atomistic resolution using the MARTINI *backward* script (see Par. 5.2 Methods).[296] The representative structure of the first minimum (M1) showed a highly symmetric interface formed by TM1-TM2 and H8 helices with contacts that span along the whole dimeric interface (Figure 5.9). Notably, in this orientation the main interaction interface is formed by the two TM1 parallel helices and from the top part (extracellular side) of TM2.

According to the cholesterol density map (Figure 5.8), no cholesterol molecules participating to dimerization were detected. In particular, in the top part of the interface (toward the extracellular side) the TM1 and TM2 helices are rich in hydrophobic amino acids, such as alanine, valine, leucine and isoleucine residues, that form stabilizing van der Waals contacts. On the other hand, in the bottom

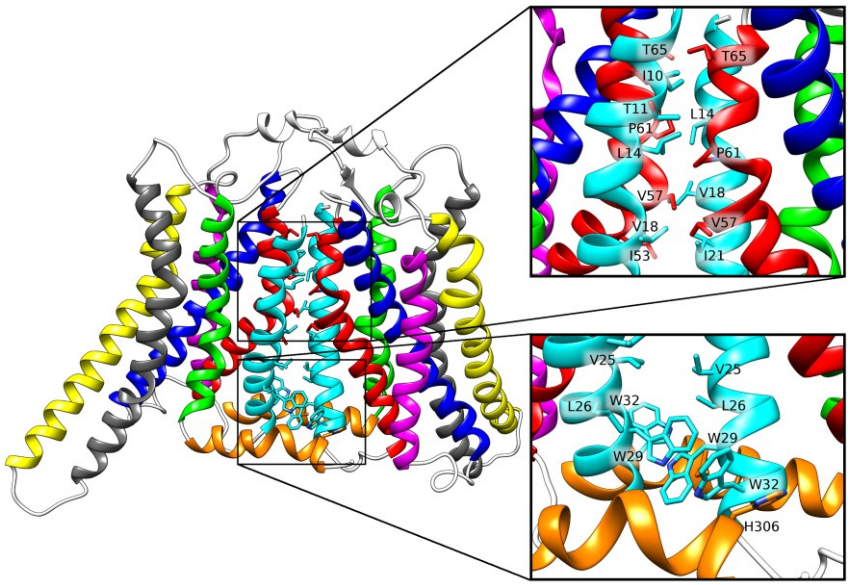

Figure 5.9: Representative conformation of the bound M1 state back-mapped to all atom resolution. Two details of the interface are shown as inset: top) the hydrophobic contacts between TM1 and TM2 helices; bottom) the interactions between W29 an W32 residues of both protomers. Protein helices are coloured according to Figure 5.1.

part of the interface (toward the intracellular side) both proteins interact mainly with their W29 and W32 residues lying on the TM1 helix. It is therefore possible to assume that this region forms a complex pi-stacking network that stabilizes the dimer conformation.

The representative structure of the second minimum (M2) presents the same orientation as the M1 conformation, but with a cholesterol molecule at the dimeric interface (Figure 5.10). A close look at the atomistic structure reveals indeed that the hydrophobic contacts on top of TM1 and TM2 are conserved with respect to the M1 conformation. However, the reciprocal contacts in the central part of the two proteins are replaced by new interactions with the cholesterol molecule. In the lower part of the interface, proteins are instead not in contact and the interaction between tryptophan residues characterizing the M1 conformation is no longer present.

The representative structure of the third minimum (M3) presents a completely different orientation. In particular, the TM1-TM2 helices are always participating in dimerization, but in this case the

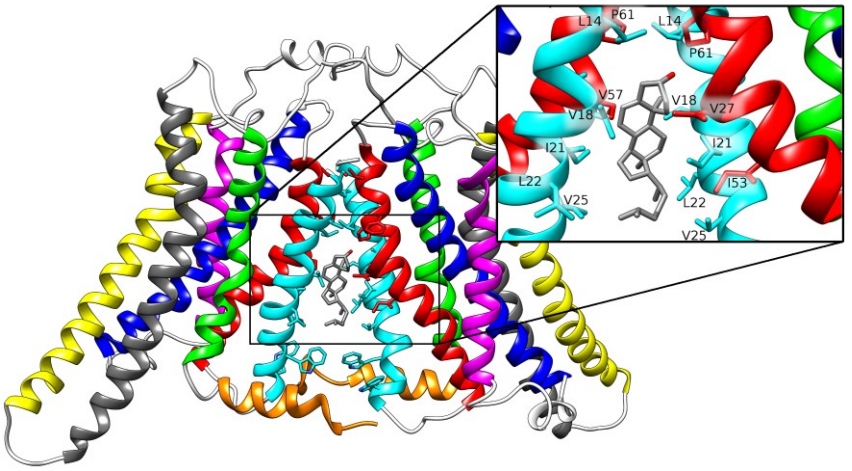

Figure 5.10: Representative conformation of the bound M2 state back-mapped to all atom resolution. Detail of the cholesterol molecule lying at the interface are showed in the right square. The colours of protein helices are the same reported in Figure 5.1, while the cholesterol molecule is represented as grey sticks.

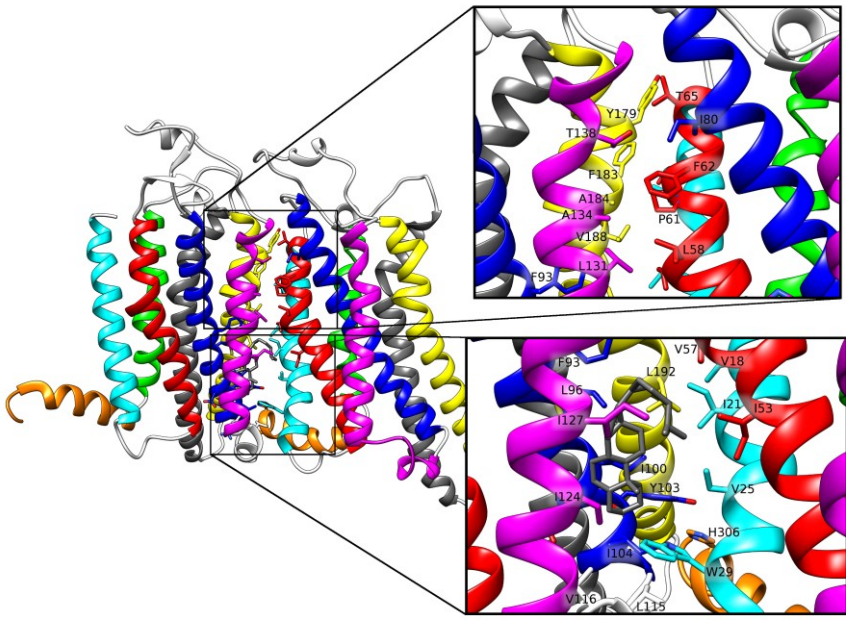

Figure 5.11: Representative conformation of the bound M3 state back-mapped to all atom resolution. Two details of the interface are shown: in the top square, hydrophobic contacts between TM1-TM2 and TM4-TM5 helices; in the bottom square, thecholesterol molecules lying at the dimeric interface. The colours of protein helices are the same reported in Figure 5.1, while the cholesterol molecule is represented as grey sticks.

orientation is not symmetric and involves mainly the TM4 and TM5 helices of the second protomer (Figure 5.11). In the top part of the dimeric interface, interactions are driven by hydrophobic contacts between residues, while in the bottom part a cholesterol molecule mediates protein-protein interactions. While in M2 the cholesterol molecule causes the loss of contacts between the protomers in the bottom part of the interface, in this case the dimer remains compact along the whole interface length. The cholesterol molecule indeed lies within a pocket formed by the TM3-TM4-TM5 helices that form a flat surface for the interaction with the other protomer. It is worth noting that the pocket that accommodates the cholesterol was already present in the unbound conformation, as can be seen in Figure 5.8,

where an increased cholesterol density is present between the TM3-TM4-TM5 helices of the unbound conformation.

In this orientation, the TM5 helix is part of the interface and is in contact with TM1 and H8 of the second protomer. Since the motion of the intracellular side of TM5-TM6 plays an essential role in the ligand-induced activation of GPCRs, allowing the binding of the G protein, it is tempting to suggest that the dimer formation can play a role in the cell signalling activated by the GPCR. To verify this hypothesis the motion of TM6 was monitored during the simulation, measuring the distance between the BB beads of residue Ile 106 and Thr 224 (Figure 5.12). In the unbound, bound M1, and bound M2 conformations, the TM6 helix is mainly in the open conformation (high distance values), but can move towards TM3 helix towards the *closed* conformation (low distance values). In the bound M3 conformation, instead, the TM5 and TM6 helices assume a more open conformation due to the presence of the H8 helix of the other protomer forming the dimer. The CG model is however constrained

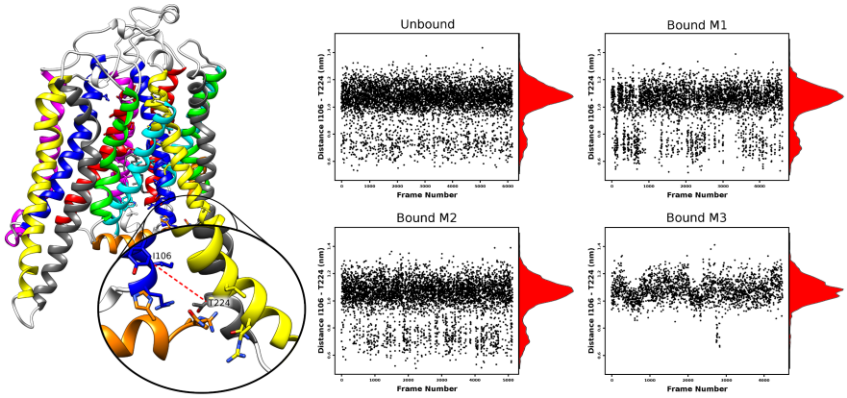

Figure 5.12: Monitoring of the distance between Ile 106 and Ala 203 BB beads. On the left: structure of the M3 bound conformation where H8 contact the TM5 helix. The distance between Ile 106 on the TM3 and Ala 203 on the TM5 is represented with red dashed line. The colours of protein helices are the same reported in Figure 5.1. On the right: plot of distance values calculated from frames of different conformations (top left: unbound, top right: bound M1; bottom left: bound M2, bottom right bound M3). A red dashed line at a distance of 0.75nm separates two states of the TM5: closed, under the line, and open, over the line.

with the ELNEDYN approach that prevent protein motions at the tertiary structure level. Further atomistic simulations starting from the dimer structure will be performed to investigate the conformational flexibility of the TM5-TM6 helix and validate the above hypothesis.

5.4 Discussion

The study of the GPCRs dimerization process is of paramount importance not only in the context of elucidation of the nature of the protein-protein interactions but also for the relevance of these systems as drugs targets for many human diseases. To contribute to such an effort, in this work the dimerization process of the A2aR GPCR protein was investigated by extensive CG-MetaD simulations.

The results obtained in our study suggest that A2aR, embedded in a membrane bilayer, mainly presents a homodimeric form, in qualitative agreement with previous FRET experiments.[291] From the analysis of the obtained FES, two different types of dimer orientation were found. The first one is a symmetric arrangement where the interaction interfaces involve the TM1-TM2-H8 element. This arrangement was already observed in Electron Microscopy (EM) and X-ray crystallography for opsin, rhodopsin, and metarhodopsins I and II.[266–268,298,299] We have disclosed two distinct free-energy minima for this dimer orientation. The first (M1) is characterized by hydrophobic contacts between the protomers on the extracellular side of the interface, and a network of pi-stacking interactions between the 29 and 32 Trp residues of each protomer. This arrangement is highly compact, and its minimum was found at a distance r of about 3.3 nm. The second minimum (M2) presents a cholesterol molecule interposed between the two protomers. The presence of the cholesterol causes the shift of the proteins and the loss of some contacts at the intracellular side, while the hydrophobic contacts at the extracellular

side are conserved. This form is less compact than M1, and its minimum was found at a distance r of about 3.8 nm.

In the second dimer orientation (M3), proteins form an asymmetric interface between the TM1-TM2-H8 element of one protomer and the TM3-TM4-TM5 of the other. As in M2, a cholesterol molecule participates at the interface, but the dimer appears highly compact with the minimum at a distance r of about 3.2 nm. The cholesterol molecule lies within a pocket between TM4-TM5 that was already present in the unbound form and that is not a specific feature of this interface. Even in this case, the involvement of the TM4-TM5 helices has been already proposed for other GPCRs.[264,300–302] In this orientation the TM5 helix participates at the interface and its internal motion is influenced by interactions with H8 of the second protomer. In particular, when the dimer is in the M3 state, TM5 and TM6 seem to be blocked in an *open* conformation, while in the other bound and unbound states they can approach the TM3 helix in a more *closed* conformation. This kind of motion is known to take part in the activation process of GPCRs: in absence of ligand, the protein is mainly found in the *closed* conformation with a *ionic lock* between residues Arg102 and Glu128; after the agonist binding, it switches to an *open* conformation breaking the ionic lock.[303] Here we have found a predominantly open state even in absence of agonist molecules. This could be due to the protein starting conformation (agonist (adenosine) bound form) and to the constrain imposed to secondary structures with the ELNEDYN approach. Even if the CG model is not suited for the study of the protein internal motions, here we hypothesize that dimerization can influence the activation process of A2aR when the asymmetric dimer (M3) is formed. Intriguingly, it was previously suggested that the μ-opioid homodimer involving the TM5-TM6 interface could preclude both protomers from properly coupling to G protein because the agonist-induced receptor–G protein interaction depends on rearrangements of TM5 and TM6.[258] Similarly, the M3 dimer orientation we have found for A2aR could influence the ability

of binding of the G protein. Moreover, due to the relevance of TM5 in both dimerization and activation process, it is important to assess if different ligands can also promote different dimeric interfaces by stabilizing different receptor conformations.

Overall, our study suggests that the TM1-TM2-H8 helices form the primary surface for dimer formation, being involved in all the dimer forms found. Our results are in good agreement with the available experimental information found for other class A GPCRs, suggesting a common dimerization pattern. Cholesterol molecules may play a key role in stabilizing these dimers, and different membrane compositions may promote different binding modes. Further theoretical investigations are required to validate the stability of the found dimer structures employing both CG and all-atom MD simulations. Our work might aid the design of experiments for a deep characterization of the A2aR dimerization process (e.g., mutagenesis and FRET experiments) and to assess the role of A2aR dimerization in cell signalling. The protocol described herein will hopefully be used to decipher the mechanistic details of other GPCRs dimerization processes with molecular details that are unattainable using current experimental techniques. To understand whether the nature of interaction is similar in other GPCR dimer forms, additional simulations on several different GPCR systems are currently ongoing.

"...we are perhaps not far removed from the time when we shall be able to submit the bulk of chemical phenomena to calculation"

Joseph Louis Gay-Lussac, -
Read before the Philomathic Society, 1808

CHAPTER 6

CONCLUSIONS

In this PhD thesis I investigated the role of the dynamic behaviour of proteins in biomolecular interactions and discussed the computational methods useful to address different problems relating to this topic. In this framework, molecular dynamics was used as a computational microscope revealing biomolecular mechanisms at spatial and temporal scales that are difficult to observe experimentally.

The ligand-binding process is of paramount importance for the development of new drugs able to alter protein functionality and the use of computational methods to support experimental research in this field is continuously growing. We selected two ligand-binding processes that represent challenging tasks due to the presence of large conformational changes of the receptor associated to the binding event (Chapter 3). Standard docking calculations would fail in treating such processes without the *a-priori* knowledge of the protein bound-state, and the inclusion of receptor flexibility is thus required. To this aim, we suggested the use of the ensemble-docking approach with multiple receptor conformations derived from molecular dynamics simulations. When conformational selection operates in the binding mechanism, proteins should be able to populate the bound conformations even in absence of ligands, thus justifying the use of simulations of the apo system to derive representative receptor conformations for docking. Conversely, when binding is characterized by conformational changes involving high-energy barriers, the operating model is the induced fit and conformational sampling of the apo receptor should be greatly extended to observe the transition. Our results suggested that ligand binding to AChBP is mainly driven by dispersive interactions that guide a conformational-selection mechanism, while the Allose BP binds ligands with strong and directed electrostatic interactions that support the involvement of induced fit. In both cases, accelerated molecular dynamics allowed sampling of protein bound conformations in absence of ligands, thus confirming the validity of the method in overcoming high-energy barriers. Moreover, the key elements in the sampling, clustering and docking stages that may lead to an effective ensemble- docking protocol emerged from our analysis, linking the mechanistic understanding of ligand binding to the development of effective computational strategies.

The role of protein dynamics in binding processes was also addressed in the case of protein-protein dimerization (Chapter 5).

Biomolecular recognition between two proteins is widely used by nature to exert specific functions, and represents an incredible opportunity to achieve functional crosstalk. In our study, the binding event was directly simulated to obtain information about the mechanism that regulates dimerization of the adenosine A2A receptor membrane protein. Due to the time- and length-scale limitations imposed by this kind of processes, we used a recently reported innovative protocol called coarse-grained metadynamics that combines the reduction of the system dimensionality through a coarse-grained description with the enhanced-sampling ability of metadynamics. The method has proved to be able to investigate the slow association and dissociation processes and identify low-energy states in the free-energy landscape. Our results indicated that the adenosine A2A receptor, embedded in a membrane bilayer, mainly presents a homodimeric form and led us to characterize the most stable reciprocal orientations of the two partners. We also found that in some of these interfaces, cholesterol molecules mediate protein-protein interactions and stabilize the dimer conformation. Future studies, with both coarse-grained and all-atom simulations starting from the obtained dimeric states, will assess the dimer stability and provide more detailed mechanistic information on the possible biological role of adenosine A2A receptor dimerization. The protocol used in this work will be used also to decipher the mechanistic details of the dimerization processes of other GPCRs, providing molecular details that are unattainable using current experimental techniques. These studies are of paramount importance not only for elucidation of the protein-protein interaction features but also for the relevance of these systems as drugs targets for many human diseases.

Finally, the ligand-induced inhibition of protein-protein dimerization was investigated (Chapter 4). Targeting protein-protein interactions with ligands is a challenging task because ligands have to be designed not only to strongly bind to the complex but also to perturb the system. In the case we investigated, the HIF-2α:ARNT

dimer structure was already known from a recent crystallographic study that also addressed the problem of ligand inhibition from a static point of view. We used conventional molecular dynamics to investigate both the determinants of protein dimerization and the perturbation induced by a ligand in the dimer stability. Residue conservation analysis and binding free-energy calculations were in good agreement in the identification of an ARNT loop relevant for protein dimerization, a feature unappreciated by the static crystal structure. Moreover, we obtained clear indications on short-range perturbations affecting a protein-protein interface nearby the binding site, as well as on an allosteric propagation of this perturbation that could promote a wider destabilization in the central and pivotal dimer interface. This last feature was derived from the observation of residue correlated motions that suggest a communication path through the dimer domains. Our results are in good agreement with the available information derived from experimental mutagenesis studies and suggest novel ligand features for successful inhibition of HIF-2α:ARNT dimerization.

The studies presented in this PhD thesis overcome the representation of binding processes with a static picture and describe the dynamic features of biomolecular interactions. This knowledge is of paramount importance for mechanistic interpretations that link binding with protein function, directing scientists toward the understanding and regulation of the complex machinery of cells.

"I am among those who think that science has great beauty. A scientist in his laboratory is not only a technician, he is also a child place before natural phenomenon, which impress him like a fairy tale."

Marie Curie -
As quoted in Madame Curie : A Biography

Appendix A

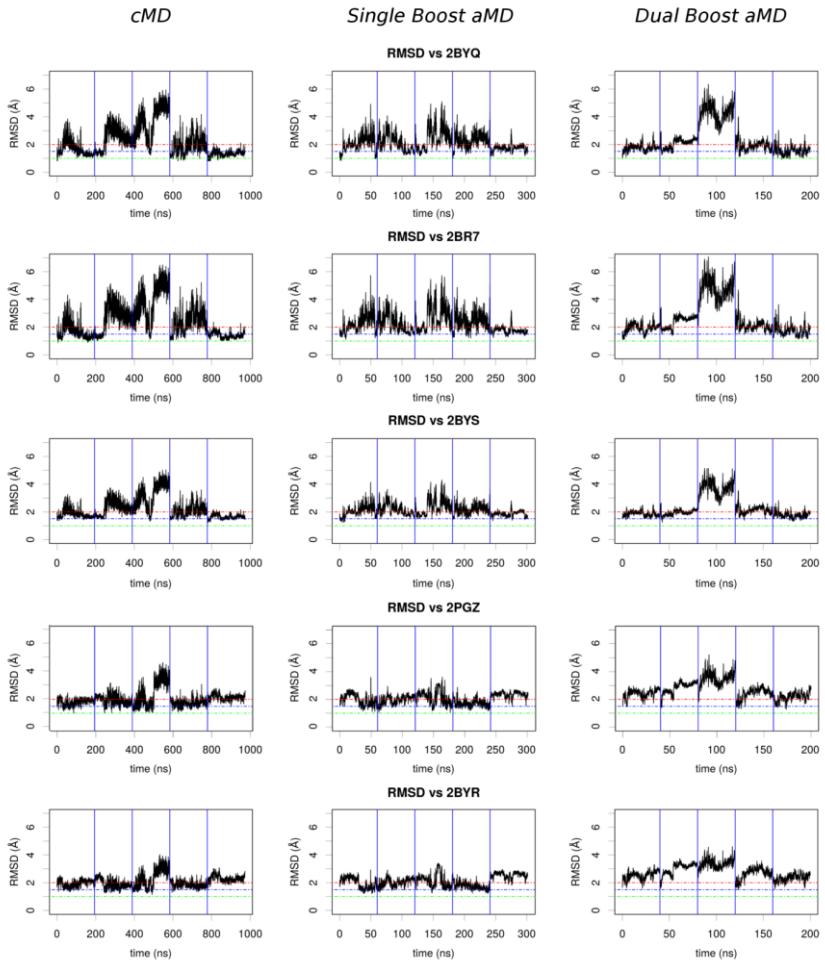

Figure App. A1: AChBP simulations. Plots of the RMSD to the 5 X-ray holo structures, calculated on heavy atoms of binding site residues, during the cMD, single boost aMD and dual boost aMD simulations.

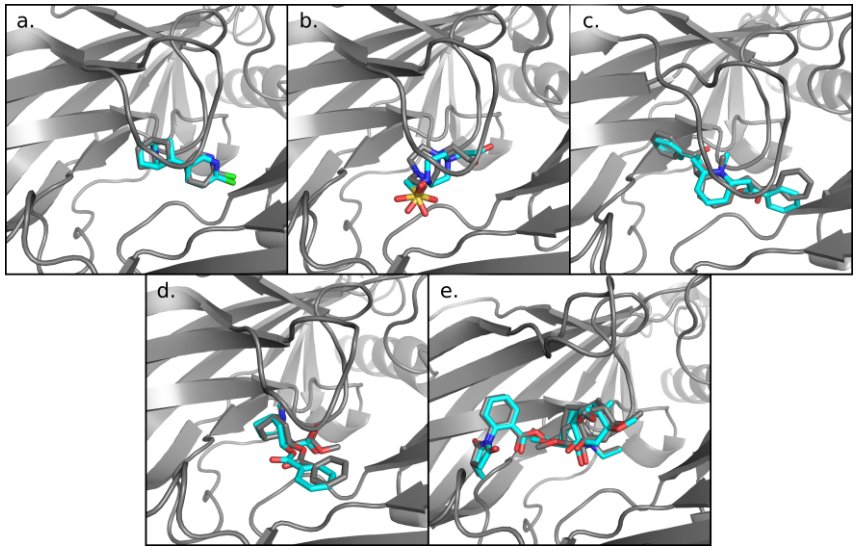

Figure App. A2: Redocking of the AChBP ligands in their own X-ray protein conformation. Docked poses are shown in cyan sticks, X-ray geometries in grey sticks. a) EPJ; b) EPE; c) LOB; d) COC; e) MLK.

Table App. A1: Length and acceleration parameters for the simulations here performed

	time (ns)	E_{plim} (kcal/mol)	α_p (kcal/mol)	E_{dlim} (kcal/mol)	α_d (kcal/mol)
		AChBP			
cMD	200				
Single Boost aMD	60	-240150	13670	-	-
Dual Boost aMD	40	-240150	13670	17644	836
		Allose BP			
cMD	1000	-	-	-	-
Single Boost aMD	1550	-147900	8000	-	-
Dual Boost aMD 1	600	-154854	10395	3579	288
Dual Boost aMD	330	-150695	8316	4155	230
Dual Boost aMD	300	-148616	5717	4443	216

Table App. A2: Redocking results for the AChBP ligands

	EPJ		EPE		LOB		COC		MLK	
	XP score	dRMSD (Å)	XP score	dRMSD (Å)	XP score	dRMSD (Å)	XP score	dRMSD (Å)	XP score	dRMSD (Å)
Pose 1	-10.94	0.29	-10.91	1.34	-15.27	0.83	-12.71	1.82	-12.56	0.33
Pose 2	-10.70	0.74	-10.30	0.88	-15.05	1.05	-12.62	2.59	-12.56	0.40
Pose 3	-5.35	1.08	-10.27	0.82	-13.03	1.08	-11.04	1.87	-11.62	0.67
Pose 4	-5.06	0.85	-10.04	1.53			-10.21	0.44		
Pose 5			-8.98	1.02						

Table App. A3: RMSD (binding site residues) to the Allose BP holo structure for representatives obtained by clustering based on the binding site residue positions.

Cluster #	Cluster size	Binding site RMSD (Å)
1	63.42%	5.64
2	13.23%	3.33
3	9.56%	4.58
4	3.34%	5.04
5	2.19%	2.71
6	1.57%	5.07
7	1.17%	4.61
8	1.05%	6.40
9	0.85%	2.99
10	0.69%	6.55
11	0.61%	4.74
12	0.46%	3.13
13	0.39%	5.09
14	0.30%	7.26
15	0.24%	1.92
16	0.19%	6.33
17	0.11%	4.25
18	0.08%	6.33
19	0.08%	6.26
20	0.05%	4.39

Appendix B

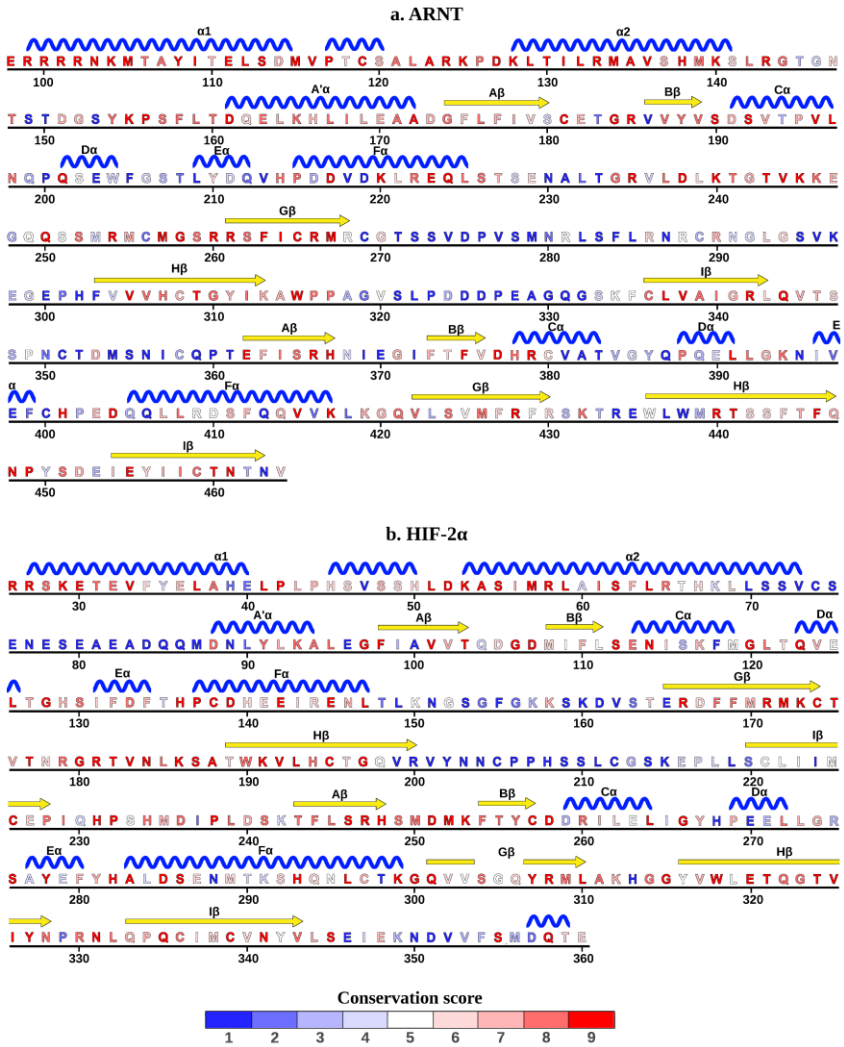

Figure App. B1: Conservation score profile (blue: low conserved, red: highly conserved) obtained by ConSurf. Secondary structure elements according to DSSP for the 4ZP4 PDB structure are reported above each sequence and labelled according to the PAS domain nomenclature.

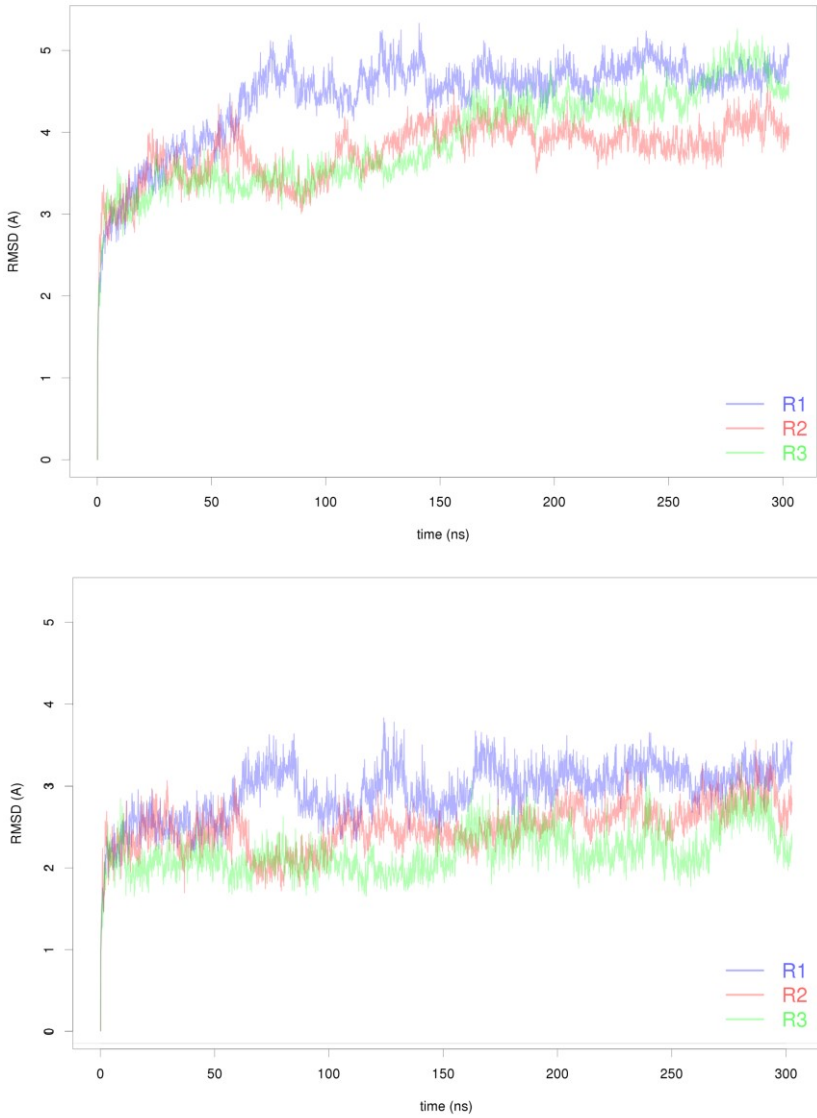

Figure App. B2: RMSD plots for the simulations of the HIF-2α:ARNT dimer in the apo form (PDB ID: 4ZP4). RMSD values are calculated on all Cα atoms (upper panel) or on the bHLH-PAS domains excluding loops and linkers (lower panel). In each panel, the RMSD for the three replicas (R1, R2, and R3) are shown.

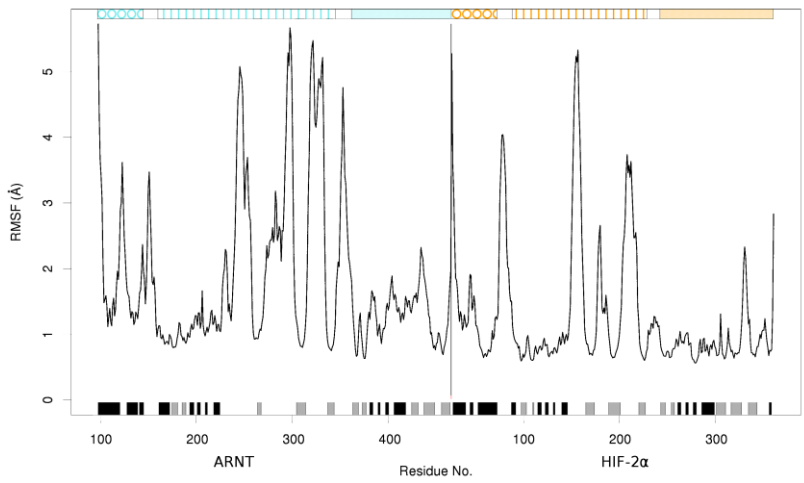

Figure App. B3: RMSF plot for the simulations of the HIF-2α:ARNT dimer in the apo form (PDB ID: 4ZP4). RMSF values are calculated on the Cα atoms. Domains are indicated on the top (ARNT: cyan; HIF-2α: orange; circle: bHLH; vertical lines: PAS-A; light filled: PAS-B) and the protein secondary structure elements according to DSSP are reported at the bottom of the graph (black: α-helix, light grey: β-strand).

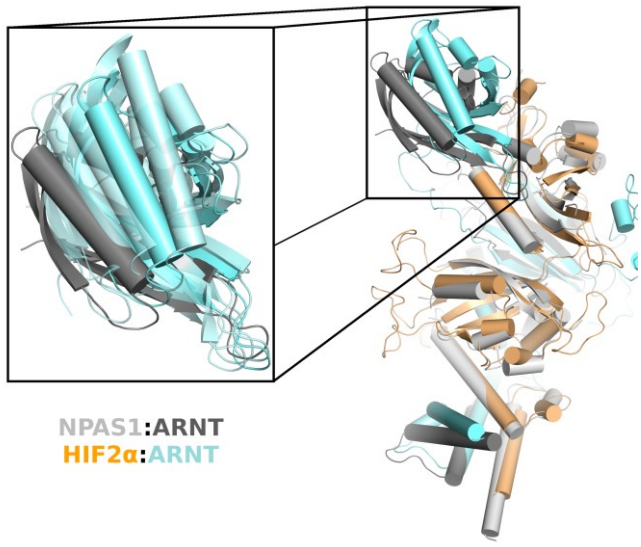

Figure App. B4: Close-up of the ARNT PAS-B structures. NPAS1:ARNT X-ray deposition (PDB 5SY5) is shown in grey; HIF-2α:ARNT X-ray deposition (PDB 4ZP4), in cyan; and three representative states extracted from MD simulations, in transparent cyan. The two complete X-ray structures are shown on the right.

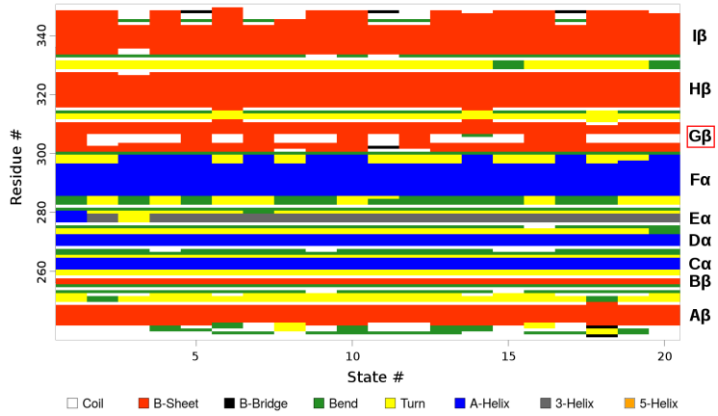

Figure App. B5: HIF-2α PAS-B secondary structures according to DSSP for the 20 conformers in the 1P97 NMR deposition. The secondary structure elements are labelled on the right according to the PAS domain nomenclature.

Table App. B1: Domain decomposition of the MM-GBSA $\Delta G_{binding}$ for the apo HIF-2α:ARNT dimer

		$\Delta G_{binding}$ (kcal mol^{-1}) (SD)[a]
	Dimer	-425,26 (0,61)
ARNT	bHLH	46,17
	Linker bHLH-PAS-A	23,86
	PAS-A	88,39
	Linker PAS-A-PAS-B	17,86
	PAS-B	34,65
HIF-2α	bHLH	47,84
	Linker bHLH-PAS-A	15,74
	PAS-A	67,43
	Linker PAS-A-PAS-B	23,19
	PAS-B	60,12

a SD: standard error of the mean, defined as $\sigma/\sqrt{n}$, where n is the number of snapshots and σ is the standard deviation between snapshots.

FIGURE DETAILS

Unless otherwise indicated, all Figures are made from scratch using the following programs:

- Pymol, Chimera or VMD for images representing 3D models of molecules;
- R for charts;
- GIMP for graphical editing.

REFERENCES

(1) Zuckerman, D. A. *Statistical Physics of Biomolecules - An Introduction*; CRC Press, Ed.; Taylor & Francis Group, 2010.

(2) Wei, G.; Xi, W.; Nussinov, R. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem. Rev.* 2016, *116*, 6516–6551.

(3) Feynman, R. P.; Leighton, R.; Sands, M. *The Feynman Lectures on Physics*; Addison–Wesley, 1964.

(4) Henzler-Wildman, K.; Kern, D. Dynamic Personalities of Proteins. *Nature* 2007, *450* (December), 964–972.

(5) 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.* 2012, *41* (1), 429–452.

(6) Kumar, S.; Ma, B.; Tsai, C.-J.; Sinha, N.; Nussinov, R. Folding and Binding Cascades: Dynamic Landscapes and Population Shifts. *Protein Sci.* 2000, *9*, 10–19.

(7) Boehr, D. D.; Nussinov, R.; Wright, P. E. The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. *Nat. Chem. Biol.* 2009, *5* (11), 789–796.

(8) Csermely, P.; Palotai, R.; Nussinov, R. Induced Fit, Conformational Selection and Independent Dynamic Segments: An Extended View of Binding Events. *Trends Biochem. Sci.* 2010, *35* (10), 539–546.

(9) Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U. S. A.* 1958, *44* (2), 98–104.

(10) Tsai, C. J.; Kumar, S.; Ma, B.; Nussinov, R. Folding Funnels, Binding Funnels, and Protein Function. *Protein Sci.* 1999, *8* (September 2000), 1181–1190.

(11) Ma, B.; Kumar, S.; Tsai, C.-J.; Nussinov, R. Folding Funnels and Binding Mechanisms. *Protein Eng.* 1999, *12* (9), 713–720.

(12) Okazaki, K.-I.; Takada, S. Dynamic Energy Landscape View of Coupled Binding and Protein Conformational Change: Induced-Fit versus

Population-Shift Mechanisms. *Proc. Natl. Acad. Sci. U. S. A.* 2008, *105* (32), 11182–11187.

(13) Changeux, J.-P.; Edelstein, S. Conformational Selection or Induced Fit? 50 Years of Debate Resolved. *F1000 Biol. Rep.* 2011, *3*, 19.

(14) Wlodarski, T.; Zagrovic, B. Conformational Selection and Induced Fit Mechanism Underlie Specificity in Noncovalent Interactions with Ubiquitin. *Proc. Natl. Acad. Sci.* 2009, *106* (46), 19346–19351.

(15) Guo, J.; Zhou, H. X. Protein Allostery and Conformational Dynamics. *Chem. Rev.* 2016, *116* (11), 6503–6515.

(16) Liu, J.; Nussinov, R. Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLoS Comput. Biol.* 2016, *12* (6), 3–7.

(17) Changeux, J. P. 50Th Anniversary of the Word "Allosteric." *Protein Sci.* 2011, *20*, 1119–1124.

(18) Nussinov, R.; Tsai, C.-J.; Ma, B. The Underappreciated Role of Allostery in the Cellular Network. *Annu. Rev. Biophys.* 2013, *42*, 169–189.

(19) Zhuravlev, P. I.; Papoian, G. A. Protein Functional Landscapes, Dynamics, Allostery: A Tortuous Path towards a Universal Theoretical Framework. *Q. Rev. Biophys.* 2010, *43*, 295–332.

(20) Nussinov, R.; Tsai, C. Allostery in Disease and in Drug Discovery. *Cell* 2013, *153* (2), 293–305.

(21) Nussinov, R.; Tsai, C. J. Unraveling Structural Mechanisms of Allosteric Drug Action. *Trends Pharmacol. Sci.* 2014, *35* (5), 256–264.

(22) Bermudez, M.; Mortier, J.; Rakers, C.; Sydow, D.; Wolber, G. More than a Look into a Crystal Ball: Protein Structure Elucidation Guided by Molecular Dynamics Simulations. *Drug Discov. Today* 2016, *21* (11), 1799–1805.

(23) Brunk, E.; Rothlisberger, U. Mixed Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States. *Chem. Rev.* 2015, *115* (12), 6217–6263.

(24) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* 1982, *161*, 269–288.

(25) Lexa, K. W.; Carlson, H. a. Protein Flexibility in Docking and Surface Mapping. *Q. Rev. Biophys.* 2012, *45* (3), 301–343.

(26) Morra, G.; Genoni, A.; Merz, K. M.; Colombo, G.; Neves, M. Molecular Recognition and Drug-Lead Identification: What Can Molecular

Simulations Tell Us? *Curr Med Chem* 2010, *17* (1), 25–41.

(27) 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.* 2004, *3* (November), 935–949.

(28) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* 2004, *432*, 862–865.

(29) Jiang, F.; Kim, S.-H. " Soft Docking ": Matching of Molecular Surface Cubes. *J. Mol. Biol.* 1991, *219*, 79–102.

(30) Totrov, M.; Abagyan, R. Flexible Ligand Docking to Multiple Receptor Conformations: A Practical Alternative. *Curr. Opin. Struct. Biol.* 2008, *18*, 178–184.

(31) Ravindranath, P. A.; Forli, S.; Goodsell, D. S.; Olson, A. J.; Sanner, M. F. AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility. *PLOS Comput. Biol.* 2015, *11* (12).

(32) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel Procedure for Moldeing Ligand/Receptor Induced Fit Effects. *J Med Chem* 2006, *49* (2), 534–553.

(33) Ding, F.; Yin, S.; Dokholyan, N. V. Rapid Flexible Docking Using a Stochastic Rotamer Library of Ligands. *J. Chem. Inf. Model.* 2010, *50* (9), 1623–1632.

(34) Davis, I. W.; Baker, D. RosettaLigand Docking with Full Ligand and Receptor Flexibility. *J. Mol. Biol.* 2009, *385* (2), 381–392.

(35) Park, H.; Lee, H.; Seok, C. High-Resolution Protein-Protein Docking by Global Optimization: Recent Advances and Future Challenges. *Curr. Opin. Struct. Biol.* 2015, *35*, 24–31.

(36) Lorenzen, S.; Zhang, Y. Monte Carlo Refinement of Rigid-Body Protein Docking Structures with Backbone Displacement and Side-Chain Optimization. *Protein Sci.* 2007, *16* (12), 2716–2725.

(37) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. Protein-Protein Docking with Simultaneous Optimization of Rigid-Body Displacement and Side-Chain Conformations. *J. Mol. Biol.* 2003, *331* (1), 281–299.

(38) Andruiser, N.; Nussinov, R.; Wolfson, H. J. FireDock: Fast Interaction Refinement in Molecular Docking. *Proteins* 2007, *69*, 139–159.

(39) Zacharias, M. Protein-Protein Docking with a Reduced Protein Model Accounting for Side-Chain Flexibility. *Protein Sci.* 2003, *12*, 1271–1282.

(40) Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J. HADDOCK : A Protein

- Protein Docking Approach Based on Biochemical or Biophysical Information. *J Am Chem Soc* 2003, *125*, 1731–1737.

(41) Mashiach, E.; Nussinov, R.; Wolfson, H. J. FiberDock: Flexible Induced-Fit Backbone Refinement in Molecular Docking. *Proteins Struct. Funct. Bioinforma.* 2010, *78* (6), 1503–1519.

(42) May, A.; Zacharias, M. Energy Minimization in Low-Frequency Normal Modes to Efficiently Allow for Global Flexibility during Systematic Protein–protein Docking. *Proteins* 2008, *70*, 794–809.

(43) Moal, I. H.; Bates, P. A. SwarmDock and the Use of Normal Modes in Protein-Protein Docking. *Int. J. Mol. Sci.* 2010, *11* (10), 3623–3648.

(44) Venkatraman, V.; Ritchie, D. W. Flexible Protein Docking Refinement Using Pose-Dependent Normal Mode Analysis. *Proteins Struct. Funct. Bioinforma.* 2012, *80* (9), 2262–2274.

(45) Pons, C.; Grosdidier, S.; Solernou, A.; Pérez-Cano, L.; Fernández-Recio, J. Present and Future Challenges and Limitations in Protein-Protein Docking. *Proteins Struct. Funct. Bioinforma.* 2010, *78* (1), 95–108.

(46) Warren, G. L.; Peishoff, C. E.; Head, M. S. Docking Algorithms and Scoring Functions; State-of-the-Art and Current Limitations. In *Computational and Structural Approaches to Drug Discovery: Ligand-Protein Interactions*; Stroud, R., Ed.; 2007; pp 137–154.

(47) Chen, Y. C. Beware of Docking! *Trends Pharmacol. Sci.* 2015, *36* (2), 78–95.

(48) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. The Role of Molecular Dynamics and Related Methods in Drug Discovery The Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* 2016, *59*, 4035–4061.

(49) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of Folded Proteins. *Nature* 1977, *267* (5612), 585–590.

(50) Levitt, M.; Warshel, A. Computer Simulation of Protein Folding. *Nature* 1975, *253*, 694–698.

(51) Feixas, F.; Lindert, S.; Sinko, W.; McCammon, J. A. Exploring the Role of Receptor Flexibility in Structure-Based Drug Discovery. *Biophys. Chem.* 2014, *186*, 31–45.

(52) Bliznyuk, A. A.; Gready, J. E. Combining Docking and Molecular Dynamic Simulations in Drug Design. *Med. Res. Rev.* 2006, *26*, 531–568.

(53) Buch, I.; Giorgino, T.; De Fabritiis, G. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations.

Proc. Natl. Acad. Sci. 2011, *108* (25), 10184–10189.

(54) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. How Does a Drug Molecule Find Its Target Binding Site? *J. Am. Chem. Soc.* 2011, *133* (24), 9181–9183.

(55) Decherchi, S.; Berteotti, A.; Bottegoni, G.; Rocchia, W.; Cavalli, A. The Ligand Binding Mechanism to Purine Nucleoside Phosphorylase Elucidated via Molecular Dynamics and Machine Learning. *Nat. Commun.* 2015, *6*, 6155.

(56) Mondal, S.; Johnston, J. M.; Wang, H.; Khelashvili, G.; Filizola, M.; Weinstein, H. Membrane Driven Spatial Organization of GPCRs. *Sci. Rep.* 2013, *3*, 1–9.

(57) Provasi, D.; Boz, M. B.; Johnston, J. M.; Filizola, M. Preferred Supramolecular Organization and Dimer Interfaces of Opioid Receptors from Simulated Self-Association. *PLoS Comput. Biol.* 2015, *11* (3), 1–21.

(58) Collier, G.; Ortiz, V. Emerging Computational Approaches for the Study of Protein Allostery. *Archives of Biochemistry and Biophysics*. Elsevier Inc. 2013, pp 6–15.

(59) Feher, V. a.; Durrant, J. D.; Van Wart, A. T.; Amaro, R. E. Computational Approaches to Mapping Allosteric Pathways. *Curr. Opin. Struct. Biol.* 2014, *25*, 98–103.

(60) Hertig, S.; Latorraca, N. R.; Dror, R. O. Revealing Atomic-Level Mechanisms of Protein Allostery with Molecular Dynamics Simulations. *PLoS Comput. Biol.* 2016, *12* (6), 1–16.

(61) McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; M.P., J. Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles. *J.Chem.Theory Comput.* 2009, *5*, 2486–2502.

(62) Pandini, A.; Fornili, A.; Fraternali, F.; Kleinjung, J. GSATools: Analysis of Allosteric Communication and Functional Local Motions Using a Structural Alphabet. *Bioinformatics* 2013, *29* (16), 2053–2055.

(63) Kong, Y.; Karplus, M. Article The Signaling Pathway of Rhodopsin. *Structure* 2007, *15*, 611–623.

(64) Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. Dynamical Networks in tRNA:protein Complexes. *Proc. Natl. Acad. Sci. U. S. A.* 2009, *106* (16), 6620–6625.

(65) Weinkam, P.; Pons, J.; Sali, A. Structure-Based Model of Allostery Predicts Coupling between Distant Sites. *Proc Natl Acad Sci U S A* 2012, *109*, 4875–4880.

(66) Vanwart, A. T.; Eargle, J.; Luthey-Schulten, Z.; Amaro, R. E. Exploring Residue Component Contributions to Dynamical Network Models of Allostery. *J. Chem. Theory Comput.* 2012, *8* (8), 2949–2961.

(67) Morra, G.; Verkhivker, G.; Colombo, G. Modeling Signal Propagation Mechanisms and Ligand- Based Conformational Dynamics of the Hsp90 Molecular Chaperone Full-Length Dimer. 2009, *5* (3), 17–21.

(68) Van Wart, A. T.; Durrant, J.; Votapka, L.; Amaro, R. E. Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis. *J. Chem. Theory Comput.* 2014, *10*, 511–517.

(69) Bernardi, R. C.; Melo, M. C. R.; Schulten, K. Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems. *Biochim. Biophys. Acta - Gen. Subj.* 2015, *1850* (5), 872–877.

(70) Sinko, W.; Lindert, S.; McCammon, J. A. Accounting for Receptor Flexibility and Enhanced Sampling Methods in Computer-Aided Drug Design. *Chem. Biol. Drug Des.* 2013, *81* (1), 41–49.

(71) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* 2016, *116* (14), 7898–7936.

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

(73) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* 1999, *314*, 141–151.

(74) Isralewitz, B.; Gao, M.; Schulten, K. Steered Molecular Dynamics and Mechanical Functions of Proteins. *Curr. Opin. Struct. Biol.* 2001, *11* (2), 224–230.

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

(76) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated Molecular Dynamics: A Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.* 2004, *120* (24), 11919–11929.

(77) 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.* 2008, *71*, 126601.

(78) Karlin, A. Emerging Structure of the Nicotinic Acetylcholine Receptors. *Nat. Rev. Neurosci.* 2002, *3* (2), 102–114.

(79) Orelle, C.; Ayvaz, T.; Everly, R. M.; Klug, C. S.; Davidson, A. L. Both Maltose-Binding Protein and ATP Are Required for Nucleotide-Binding

Domain Closure in the Intact Maltose ABC Transporter. *Proc Natl Acad Sci U S A* 2008, *105* (35), 12837–12842.

(80) Felder, C. B.; Graul, R. C.; Lee, A. Y.; Merkle, H.-P.; Sadee, W. The Venus Flytrap of Periplasmic Binding Proteins: An Ancient Protein Module Present in Multiple Drug Receptors. *AAPS PharmSci* 1999, *1* (2), 7–26.

(81) Motta, S.; Bonati, L. Modeling Binding with Large Conformational Changes: Key Points in Ensemble-Docking Approaches. *J. Chem. Inf. Model.* 2017, *57* (7), 1563–1578.

(82) 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.* 2009, *106* (2), 450–455.

(83) 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.* 2013, *56*, 1739–1747.

(84) 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.* 2009, *131* (48), 17647–17654.

(85) Wu, D.; Potluri, N.; Lu, J.; Kim, Y.; Rastinejad, F. Structural Integration in Hypoxia-Inducible Factors. *Nature* 2015, *524*, 303–309.

(86) Motta, S.; Minici, C.; Corrada, D.; Bonati, L.; Pandini, A. Molecular Dynamics of HIF-2α:ARNT Ligand-Induced Inhibition. *PLoS Comput. Biol. - Accepted,* 10.1371/journal.pcbi.1006021.

(87) Adcock, S. A.; McCammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* 2006, *106* (5), 1589–1615.

(88) Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* 2002, *9* (9), 646–652.

(89) Leach, A. R. Empirical Force Field Models: Molecular Mechanics. In *Molecular Modelling*; Pearson education, 2001; pp 165–252.

(90) Abraham, M. J.; van der Spoel, D.; Lindahl, E.; Hess, B.; Development, G. T. Interaction Function and Force Fields. In *Gromacs user manual Version 2016*; 2016; pp 65–116.

(91) Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham III, T. E.; Darden, T. A.; Duke, R. E.; Gohlke, H.;

Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Kollman, X.; P.A., W. Reading and Modifying Amber Parameter Files. In *Amber 14 Reference Manual*; 2014; pp 225–254.

(92) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* 1993, *97*, 10269–10280.

(93) Leach, A. R. Energy Minimization and Related Methods for Exploring the Energy Surface. In *Molecular modelling principles and application*; Pearson education, 2001; pp 253–301.

(94) Leach, A. R. Molecular Dynamics Simulation Methods. In *Molecular modelling principles and application*; Pearson education, 2001; pp 353–409.

(95) Newton, I. *Philosophiae Naturalis Principia Mathematica*; London, 1687.

(96) Verlet, L. Computers Experiments on Classical Fluids. I. Thermodynamical Properties of Lennard - Jones Molecules. *Phys. Rev.* 1967, *159* (1), 98–103.

(97) Hockney, R. W. The Potential Calculations and Some Applications. *Methods Comput. Phys.* 1970, *9* (9), 136–211.

(98) 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. Comp. Phys.* 1977, *23*, 327–341.

(99) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS : A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* 1997, *18* (12), 1463–1472.

(100) Barker, J. A.; Watts, R. O. Monte Carlo Studies of the Dielectric Properties of Water-like Models. *Mol. Phys.* 1973, *26*, 789–792.

(101) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N•log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* 1993, *98* (12), 10089–10092.

(102) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* 1995, *103* (November), 8577–8593.

(103) 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.* 1984, *81* (8), 3684–3690.

(104) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity-Rescaling. *J. Chem. Phys.* 2007, *126*, 14101.

(105) Chandler, D. *Introduction to Modern Statistical Mechaincs*; Oxford University Press, 1987.

(106) Van Gunsteren, W. F.; Daura, X.; Mark, A. E. Computation of Free Energy. *Helv. Chim. Acta* 2002, *85* (10), 3113–3129.

(107) Takada, S. Coarse-Grained Molecular Simulations of Large Biomolecules. *Curr. Opin. Struct. Biol.* 2012, *22* (2), 130–137.

(108) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B* 2004, *108* (2), 750–760.

(109) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; De Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* 2007, *111* (27), 7812–7824.

(110) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI Coarse Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* 2008, *4* (5), 819–834.

(111) De Jong, D. H.; Singh, G.; Bennett, W. F. D.; Arnarez, C.; Wassenaar, T. A.; Schäfer, L. V.; Periole, X.; Tieleman, D. P.; Marrink, S. J. Improved Parameters for the Martini Coarse-Grained Protein Force Field. *J. Chem. Theory Comput.* 2013, *9* (1), 687–697.

(112) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini Model. *Chem. Soc. Rev.* 2013, *42* (16), 6801.

(113) Periole, X.; Cavalli, M.; Marrink, S.-J.; Ceruso, M. A. Combining an Elastic Network with a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition. *J Chem Theory Comput* 2009, *5* (9), 1–7.

(114) Fritz, D.; Koschke, K.; Harmandaris, V. A.; van der Vegt, N. F. A.; Kremer, K. Multiscale Modeling of Soft Matter: Scaling of Dynamics. *Phys. Chem. Chem. Phys.* 2011, *13* (22), 10412.

(115) Doshi, U.; Hamelberg, D. Towards Fast, Rigorous and Efficient Conformational Sampling of Biomolecules: Advances in Accelerated Molecular Dynamics. *Biochim. Biophys. Acta J.* 2015, *1850*, 878–888.

(116) Miao, Y.; Nichols, S. E.; McCammon, J. A. Free Energy Landscape of G-Protein Coupled Receptors, Explored by Accelerated Molecular Dynamics. *Phys. Chem. Chem. Phys.* 2014, *16* (14), 6398.

(117) Markwick, P. R. L.; McCammon, J. A. Studying Functional Dynamics in Bio-Molecules Using Accelerated Molecular Dynamics. *Phys. Chem. Chem. Phys.* 2011, *13* (45), 20053.

(118) de Oliveira, C. A.; Hamelberg, D.; McCammon, J. A. On the Application of Accelerated Molecular Dynamics to Liquid Water Simulations. *J. Phys. Chem. B* 2006, *110* (45), 22695–22701.

(119) Hamelberg, D.; De Oliveira, C. A. F.; McCammon, J. A. Sampling of Slow Diffusive Conformational Transitions with Accelerated Molecular Dynamics. *J. Chem. Phys.* 2007, *127*, 155102.

(120) Miao, Y.; Sinko, W.; Pierce, L.; Bucher, D.; Walker, R. C.; Mccammon, J. A. Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation. *J. Chem. Theory Comput.* 2014, *10*, 2677–2689.

(121) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* 2002, *99*, 12562–12566.

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

(123) Micheletti, C.; Laio, A.; Parrinello, M. Reconstructing the Density of States by History-Dependent Metadynamics. *Phys. Rev. Lett.* 2004, *92* (17), 170601–1.

(124) Crespo, Y.; Marinelli, F.; Pietrucci, F.; Laio, A. Metadynamics Convergence Law in a Multidimensional System. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* 2010, *81* (5), 1–4.

(125) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* 2008, *100* (January), 1–4.

(126) Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. Efficient Reconstruction of Complex Free Energy Landscapes by Multiple Walkers Metadynamics. *J. Phys. Chem. B* 2006, *110* (8), 3533–3539.

(127) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. Free-Energy Landscape for β Hairpin Folding from Combined Parallel Tempering and Metadynamics. *J. Am. Chem. Soc.* 2006, *128* (41), 13435–13441.

(128) Piana, S.; Laio, A. A Bias-Exchange Approach to Protein Folding. *J. Phys. Chem. B* 2007, *111* (17), 4553–4559.

(129) Bussi, G.; Laio, A.; Parrinello, M. Equilibrium Free Energies from Nonequilibrium Metadynamics. *Phys. Rev. Lett.* 2006, *96*, 90601.

(130) Truhlar, D. G. Valence Bond Theory for Chemical Dynamics. *J. Comput.*

Chem. 2009, *28* (1), 73–86.

(131) Tiwary, P.; Parrinello, M. A Time-Independent Free Energy Estimator for Metadynamics. *J. Phys. Chem. B* 2015, *119* (3), 736–742.

(132) Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. Target Flexibility: An Emerging Consideration in Drug Discovery and Design. *J Med Chem* 2008, *51* (20), 6237–6255.

(133) Lill, M. A. Efficient Incorporation of Protein Flexibility and Dynamics into Molecular Docking Simulations. *Biochemistry* 2011, *50* (28), 6157–6169.

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

(135) Amaro, R. E.; Li, W. W. Emerging Methods for Ensemble-Based Virtual Screening. *Curr. Top. Med. Chem.* 2010, *10*, 3–13.

(136) Lin, J.-H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. Computational Drug Design Accommodating Receptor Flexibility: The Relaxed Complex Scheme. *J. Am. Chem. Soc.* 2002, *124* (20), 5632–5633.

(137) Lin, J.-H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. The Relaxed Complex Method: Accommodating Receptor Flexibility for Drug Design with an Improved Scoring Scheme. *Biopolymers* 2003, *68* (1), 47–62.

(138) Bolia, A.; Gerek, Z. N.; Ozkan, S. B. BP-Dock: A Flexible Docking Scheme for Exploring Protein-Ligand Interactions Based on Unbound Structures. *J. Chem. Inf. Model.* 2014, *54* (3), 913–925.

(139) Rueda, M.; Totrov, M.; Abagyan, R. ALiBERO: Evolving a Team of Complementary Pocket Conformations rather than a Single Leader. *J. Chem. Inf. Model.* 2012, *52* (10), 2705–2714.

(140) Karplus, M.; Kuriyan, J. Molecular Dynamics and Protein Function. *Proc. Natl. Acad. Sci.* 2005, *102* (19), 6679–6685.

(141) Van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glättli, A.; Hünenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; Van Der Vegt, N. F. A.; Yu, H. B. Biomolecular Modeling: Goals, Problems, Perspectives. *Angew. Chemie - Int. Ed.* 2006, *45* (25), 4064–4092.

(142) Maximova, T.; Moffatt, R.; Ma, B.; Nussinov, R.; Shehu, A. Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics. *PLoS Comput. Biol.* 2016, *12* (4), e1004619.

(143) Sinko, W.; de Oliveira, C.; Williams, S.; Van Wynsberghe, A.; Durrant, J. D.; Cao, R.; Oldfield, E.; Mccammon, J. A. Applying Molecular Dynamics Simulations to Identify Rarely Sampled Ligand-Bound Conformational States of Undecaprenyl Pyrophosphate Synthase, an Antibacterial Target. *Chem. Biol. Drug Des.* 2011, *77*, 412–420.

(144) Hamelberg, D.; De Oliveira, C. A. F.; McCammon, J. A. Sampling of Slow Diffusive Conformational Transitions with Accelerated Molecular Dynamics. *J. Chem. Phys.* 2007, *127*, 155120.

(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.* 2008, *22*, 693–705.

(146) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. Clusteing Molecular Dynamic Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theory Comput.* 2007, *3*, 2312–2334.

(147) 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.* 2013, *13* (5), 642–651.

(148) Hansen, S. B.; Sulzenbacher, G.; Huxford, T.; Marchot, P.; Taylor, P.; Bourne, Y. Structures of Aplysia AChBP Complexes with Nicotinic Agonists and Antagonists Reveal Distinctive Binding Interfaces and Conformations. *EMBO J.* 2005, *24* (September), 3635–3646.

(149) Celie, P. H. N.; Kasheverov, I. E.; Mordvintsev, D. Y.; Hogg, R. C.; van Nierop, P.; van Elk, R.; van Rossum-Fikkert, S. E.; Zhmak, M. N.; Bertrand, D.; Tsetlin, V.; Sixma, T. K.; Smit, A. B. Crystal Structure of Nicotinic Acetylcholine Receptor Homolog AChBP in Complex with an Alpha-Conotoxin PnIA Variant. *Nat. Struct. Mol. Biol.* 2005, *12* (7), 582–588.

(150) Hansen, S. B.; Taylor, P. Galanthamine and Non-Competitive Inhibitor Binding to ACh-Binding Protein: Evidence for a Binding Site on Non-α-Subunit Interfaces of Heteromeric Neuronal Nicotinic Receptors. *J. Mol. Biol.* 2007, *369* (4), 895–901.

(151) Chaudhuri, B. N.; Ko, J.; Park, C.; Jones, T. A.; Mowbray, S. L. Structure of D -Allose Binding Protein from Escherichia Coli Bound to D -Allose at 1.8 Å Resolution. *J. Mol. Biol.* 1999, *286* (5), 1519–1531.

(152) Magnusson, U.; Chaudhuri, B. N.; Ko, J.; Park, C.; Jones, T. A.;

Mowbray, S. L. Hinge-Bending Motion of D-Allose-Binding Protein from Escherichia Coli. Three Open Conformations. *J. Biol. Chem.* 2002, *277* (16), 14077–14084.

(153) Tam, R.; Saier Jr, M. H. Structural, Functional, and Evolutionary Relationships among Extracellular Solute-Binding Receptors of Bacteria. *Microbiol. Rev.* 1993, *57* (2), 320–346.

(154) 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.* 2000, *28*, 235–242.

(155) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W.; Sastry, G.M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *J. Comput. Aid. Mol. Des* 2013, *27* (3), 221–234.

(156) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very Fast Prediction and Rationalization of pKa Values for Protein-Ligand Complexes. *Proteins Struct. Funct. Genet.* 2008, *73* (3), 765–783.

(157) Case, D. a; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* 2005, *26* (16), 1668–1688.

(158) Salomon-Ferrer, R.; Case, D. a.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2013, *3* (2), 198–210.

(159) 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.* 2015, *11*, 3696–3713.

(160) 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.* 1983, *79* (2), 926–935.

(161) 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.* 2006, *25* (2), 247–260.

(162) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. a; Case, D. a. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* 2004, *25* (9), 1157–1174.

(163) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I.

Method. *J. Comput. Chem.* 2000, *21* (2), 132–146.

(164) 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.* 2002, *23* (16), 1623–1641.

(165) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. Langevin Dynamics of Peptides: The Frictional Dependence of Isomerization Rates of N-Acetylalanyl-N'-methylamide. *Biopolymers* 1992, *32* (5), 523–535.

(166) Pierce, L. C. T.; Salomon-Ferrer, R.; de Oliveira, C. A. F.; McCammon, J. A.; Walker, R. C. Routine Access to Millsecond Timescale Evenrs with Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* 2012, *8*, 2997–3002.

(167) Miao, Y.; Nichols, S. E.; McCammon, J. A. Free Energy Landscape of G-Protein Coupled Receptors, Explored by Accelerated Molecular Dynamics. *Phys. Chem. Chem. Phys.* 2014, *16* (14), 6398–6406.

(168) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide Folding : When Simulation Meets Experiment. *Angew. Chem. Int. Ed* 1999, *111*, 236–240.

(169) Durrant, J. D.; De Oliveira, C. A. F.; McCammon, J. A. POVME: An Algorithm for Measuring Binding-Pocket Volumes. *J Mol Graph Model* 2011, *29* (5), 773–776.

(170) Durrant, J. D.; Votapka, L.; Sørensen, J.; Amaro, R. E. POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J. Chem. Theory Comput.* 2014, *10*, 5047–5056.

(171) Swift, R. V.; Jusoh, S. A.; Offutt, T. L.; Li, E. S.; Amaro, R. E. Knowledge-Based Methods to Train and Optimize Virtual Screening Ensembles. *J. Chem. Inf. Model.* 2016, *56*, 830–842.

(172) Schrödinger, L. Schrödinger Release 2014-1: Maestro, Version 9.7. New York, NY 2014.

(173) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* 2004, *47* (7), 1739–1749.

(174) 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.* 2006, *49* (21), 6177–6196.

(175) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* 2004, *47* (7), 1750–1759.

(176) Bordogna, A.; Pandini, A.; Bonati, L. Predicting the Accuracy of Protein-Ligand Docking on Homology Models. *J. Comput. Chem.* 2011, *32* (1), 81–98.

(177) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* 1996, *14* (1), 33–38.

(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* 2006, *22* (21), 2695–2696.

(179) Skjærven, L.; Yao, X.-Q.; Scarabelli, G.; Grant, B. J. Integrating Protein Structural Dynamics and Evolutionary Analysis with Bio3D. *BMC Bioinformatics* 2014, *15* (1), 399–409.

(180) L.L.C. Schrodinger. The PyMOL Molecular Graphics System. 2010.

(181) Babakhani, A.; Talley, T. T.; Taylor, P.; McCammon, J. A. A Virtual Screening Study of the Acetylcholine Binding Protein Using a Relaxed-Complex Approach. *Comput. Biol. Chem.* 2009, *33*, 160–170.

(182) Ma, B.; Shatsky, M.; Wolfson, H. J.; Nussinov, R. Multiple Diverse Ligands Binding at a Single Protein Site: A Matter of Pre-Existing Populations. *Protein Sci.* 2002, *11* (2), 184–197.

(183) Mao, B.; Pear, M.; McCammon, J.; Quiocho, F. Hinge-Bending in L-Arabinose-Binding Protein. The "Venus's-Flytrap" model. *J. Biol. Chem.* 1982, *257* (3), 1131–1133.

(184) Bucher, D.; Grant, B. J.; Markwick, P. R.; McCammon, J. A. Accessing a Hidden Conformation of the Maltose Binding Protein Using Accelerated Molecular Dynamics. *PLoS Comput. Biol.* 2011, *7* (4).

(185) Bucher, D.; Grant, B. J.; McCammon, J. A. Induced Fit or Conformational Selection? The Role of the Semi-Closed State in the Maltose Binding Protein. *Biochemistry* 2011, *50*, 10530–10539.

(186) Tang, C.; Schwieters, C. D.; Clore, G. M. Open-to-Closed Transition in Apo Maltose-Binding Protein Observed by Paramagnetic NMR. *Nature* 2007, *449* (7165), 1078–1082.

(187) Keller, B.; Daura, X.; Van Gunsteren, W. F. Comparing Geometric and Kinetic Cluster Algorithms for Molecular Simulation Data. *J. Chem. Phys.* 2010, *132* (7), 74110.

(188) Lee, H. S.; Choi, J.; Kufareva, I.; Abagyan, R.; Filikov, A.; Yang, Y.;

Yoon, S. Optimization of High Throughput Virtual Screening by Combining Shape-Matching and Docking Methods. *J. Chem. Inf. Model.* 2008, *48* (3), 489–497.

(189) Warren, G. L.; Andrews, C. V; Capelli, a; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring functions.Capelli, A.LaLonde, J.Semus, S. F.Head, M. S. *J. Med. Chem.* 2006, No. 49, 5912–5931.

(190) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors. *J. Chem. Inf. Model.* 2006, *46* (1), 401–415.

(191) Bersten, D. C.; Sullivan, A. E.; Peet, D. J.; Whitelaw, M. L. bHLH-PAS Proteins in Cancer. *Nat. Rev. Cancer* 2013, *13* (12), 827–841.

(192) 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.* 2004, *36* (2), 189–204.

(193) Wu, D.; Rastinejad, F. Structural Characterization of Mammalian bHLH-PAS Transcription Factors. *Curr. Opin. Struct. Biol.* 2017, *43*, 1–9.

(194) Hu, C.-J.; Wang, L.-Y.; Chodosh, L. A.; Keith, B.; Simon, M. C. Differential Roles of Hypoxia-Inducible Factor 1alpha (HIF-1alpha) and HIF-2alpha in Hypoxic Gene Regulation. *Mol. Cell. Biol.* 2003, *23* (24), 9361–9374.

(195) Pongratz, I.; Antonsson, C.; Whitelaw, M. L.; Poellinger, L. Role of the PAS Domain in Regulation of Dimerization and DNA Binding Specificity of the Dioxin Receptor. *Mol. Cell. Biol.* 1998, *18* (7), 4079–4088.

(196) Crews, S. T.; Fan, C.-M. Remembrance of Things PAS: Regulation of Development by bHLH–PAS Proteins. *Curr. Opin. Genet. Dev.* 1999, *9*, 580–587.

(197) Taylor, B. L.; Zhulin, I. B. PAS Domains: Internal Sensors of Oxygen, Redox Potential, and Light. *Microbiol. Mol. Biol. Rev.* 1999, *63* (2), 479–506.

(198) Harper, S. M.; Neil, L. C.; Gardner, K. H. Structural Basis of a Phototropin Light Switch. *Science* 2003, *301* (5639), 1541–1544.

(199) Dewhirst, M. W.; Cao, Y.; Moeller, B. Cycling Hypoxia and Free Radicals Regulate Angiogenesis and Radiotherapy Response. *Nat. Rev. Cancer* 2008, *8* (6), 425–437.

(200) Kaelin, W. G.; Jr. Cancer and Altered Metabolism: Potential Importance of Hypoxia-Inducible Factor and 2-Oxoglutarate-Dependent Dioxygenases. *Cold Spring Harb. Symp. Quant. Biol.* 2011, *76*, 335–345.

(201) Keith, B.; Johnson, R. S.; Simon, M. C. HIF1α and HIF2α: Sibling Rivalry in Hypoxic Tumour Growth and Progression. *Nat. Rev. Cancer* 2011, *12* (1), 9–22.

(202) Semenza, G. L. Defining the Role of Hypoxia-Inducible Factor 1 in Cancer Biology and Therapeutics. *Oncogene* 2010, *29* (5), 625–634.

(203) Li, L.; Zhang, L.; Zhang, X.; Yan, Q.; Minamishima, Y. A.; Olumi, A. F.; Mao, M.; Bartz, S.; Kaelin, W. G. Hypoxia-Inducible Factor Linked to Differential Kidney Cancer Risk Seen with Type 2A and Type 2B VHL Mutations. *Mol. Cell. Biol.* 2007, *27* (15), 5381–5392.

(204) Morris, M. R.; Hughes, D. J.; Tian, Y.-M.; Ricketts, C. J.; Lau, K. W.; Gentle, D.; Shuib, S.; Serrano-Fernandez, P.; Lubinski, J.; Wiesener, M. S.; Pugh, C. W.; Latif, F.; Ratcliffe, P. J.; Maher, E. R. Mutation Analysis of Hypoxia-Inducible Factors HIF1A and HIF2A in Renal Cell Carcinoma. *Anticancer Res.* 2009, *29* (11), 4337–4343.

(205) Wu, D.; Su, X.; Potluri, N.; Kim, Y.; Rastinejad, F. NPAS1-ARNT and NPAS3-ARNT Crystal Structures Implicate the bHLH-PAS Family as Multi-Ligand Binding Transcription Factors. *Elife* 2016, *5*, 1–15.

(206) 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.* 2011, *124* (1), 1–22.

(207) Bonati, L.; Corrada, D.; Giani Tagliabue, S.; Motta, S. Molecular Modeling of the AhR Structure and Interactions Can Shed Light on Ligand-Dependent Activation and Transformation Mechanisms. *Curr. Opin. Toxicol.* 2017, *2*, 42–49.

(208) Halavaty, A. S.; Moffat, K. N- and C-Terminal Flanking Regions Modulate Light-Induced Signal Transduction in the LOV2 Domain of the Blue Light Sensor Phototropin 1 from Avena Sativa. *Biochemistry* 2007, *46* (49), 14001–14009.

(209) Möglich, A.; Ayers, R. A.; Moffat, K. Structure and Signaling Mechanism of Per-ARNT-Sim Domains. *Structure* 2009, *17* (10), 1282–1294.

(210) Koehler, A. N. A Complex Task? Direct Modulation of Transcription Factors with Small Molecules. *Curr. Opin. Chem. Biol.* 2010, *14* (3), 331–340.

(211) Lee, K.; Zhang, H.; Qian, D. Z.; Rey, S.; Liu, J. O.; Semenza, G. L. Acriflavine Inhibits HIF-1 Dimerization, Tumor Growth, and

Vascularization. *Proc. Natl. Acad. Sci. U. S. A.* 2009, *106* (42), 17910–17915.

(212) 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.* 2015, *58*, 5930–5941.

(213) 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.* 2013, *9* (4), 271–276.

(214) Homeyer, N.; Gohlke, H. Free Energy Calculations by the Molecular Mechanics Poisson−Boltzmann Surface Area Method. *Mol. Inform.* 2012, *31* (2), 114–122.

(215) Gohlke, H.; Kiel, C.; Case, D. A. Insights into Protein–Protein Binding by Binding Free Energy Calculation and Free Energy Decomposition for the Ras–Raf and Ras–RalGDS Complexes. *J. Mol. Biol.* 2003, *330* (4), 891–913.

(216) Venken, T.; Daelemans, D.; Maeyer, M. De; Voet, A. Computational Investigation of the HIV-1 Rev Multimerization Using Molecular Dynamics Simulations and Binding Free Energy Calculations. *Proteins* 2012, *80*, 1633–1646.

(217) Corrada, D.; Denison, M. S.; Bonati, L. Structural Modeling of the AhR:ARNT Complex in the bHLH–PASA–PASB Region Elucidates the Key Determinants of Dimerization. *Mol. BioSyst.* 2017, *13*, 981–990.

(218) Mandell, D. J.; Coutsias, E. A.; Kortemme, T. Sub-Angstrom Accuracy in Protein Loop Reconstruction by Robotics-Inspired Conformational Sampling. *Nat. Methods* 2009, *6* (8), 551–552.

(219) Stein, A.; Kortemme, T. Improvements to Robotics-Inspired Conformational Sampling in Rosetta. *PLoS One* 2013, *8* (5), e63090.

(220) Conchúir, S. Ó.; Barlow, K. A.; Pache, R. A.; Ollikainen, N.; Kundert, K.; Meara, M. J. O.; Smith, C. A.; Kortemme, T. A Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling and Design. *PLoS One* 2015, *10*, e0130433.

(221) Fraccalvieri, D.; Pandini, A.; Stella, F.; Bonati, L. Conformational and Functional Analysis of Molecular Dynamics Trajectories by Self-Organising Maps. *BMC Bioinformatics* 2011, *12* (1), 158.

(222) 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.* 2012, *8* (10), 2680.

(223) Fiser, A.; Kinh Gian Do, R.; Sali. Modeling Loops in Protein Structures. *Protein Sci.* 2000, *9*, 1753–1773.

(224) 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. *SoftwareX* 2015, *1–2*, 19–25.

(225) Aliev, A. E.; Kulke, M.; Khaneja, H. S.; Chudasama, V.; Sheppard, T. D.; Lanigan, R. M. Motional Timescale Predictions by Molecular Dynamics Simulations: Case Study Using Proline and Hydroxyproline Sidechain Dynamics. *Proteins Struct. Funct. Bioinforma.* 2014, *82* (2), 195–215.

(226) Fornili, A.; Pandini, A.; Lu, H.-C.; Fraternali, F. Specialized Dynamical Properties of Promiscuous Residues Revealed by Simulated Conformational Ensembles. *J. Chem. Theory Comput.* 2013, *9* (11), 5127–5147.

(227) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals : A New Molecular Dynamics Method Polymorphic Transitions in Single Crystals : A New Molecular Dynamics Method. *J. Appl. Phys.* 1981, *52*, 7182–7190.

(228) R-Development-Core-Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria 2014.

(229) Kabsch, W.; Sander, C. Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 1983, *22*, 2577–2637.

(230) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.* 2000, *33* (12), 889–897.

(231) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate−DNA Helices. *J. Am. Chem. Soc.* 1998, *120* (37), 9401–9409.

(232) Miller, B. R.; Mcgee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA . Py : An E Ffi Cient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* 2012, *8*, 3314–3321.

(233) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium. *J. Phys. Chem.* 1996, *100* (51), 19824–19839.

(234) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins Struct. Funct. Bioinforma.* 2004, *55* (2), 383–394.

(235) Weiser, J.; Shenkin, P. S.; Still, W. C. Approximate Atomic Surfaces from Linear Combinations of Pairwise Overlaps (LCPO). *J. Comput. Chem.* 1999, *20* (2), 217–230.

(236) Scarabelli, G.; Grant, B. J. Mapping the Structural and Dynamical Features of Kinesin Motor Domains. *PLoS Comput. Biol.* 2013, *9* (11).

(237) Goldenberg, O.; Erez, E.; Nimrod, G.; Ben-Tal, N. The ConSurf-DB: Pre-Calculated Evolutionary Conservation Profiles of Protein Structures. *Nucleic Acids Res.* 2009, *37* (SUPPL. 1), 323–327.

(238) Celniker, G.; Nimrod, G.; Ashkenazy, H.; Glaser, F.; Martz, E.; Mayrose, I.; Pupko, T.; Ben-Tal, N. ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. *Isr. J. Chem.* 2013, *53* (3–4), 199–206.

(239) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST : A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* 1997, *25* (17), 3389–3402.

(240) The Uniprot Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2008, *36*, 190–195.

(241) Edgar, R. C. MUSCLE : A Multiple Sequence Alignment Method with Reduced Time and Space Complexity. *BMC Bioinformatics* 2004, *5*, 1–19.

(242) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Soding, J.; Thompson, J. D.; Higgins, D. G. Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Mol. Syst. Biol.* 2014, *7* (1), 539–539.

(243) Robert, X.; Gouet, P. Deciphering Key Features in Protein Structures with the New ENDscript Server. *Nucleic Acids Res.* 2014, *42* (Web Server issue), W320-4.

(244) Armon, A.; Graur, D.; Ben-Tal, N. ConSurf: An Algorithmic Tool for the Identification of Functional Regions in Proteins by Surface Mapping of Phylogenetic Information. *J. Mol. Biol.* 2001, *307* (1), 447–463.

(245) Scarabelli, G.; Grant, B. J. Kinesin-5 Allosteric Inhibitors Uncouple the Dynamics of Nucleotide, Microtubule, and Neck-Linker Binding Sites. *Biophys. J.* 2014, *107* (9), 2204–2213.

(246) 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.* 2016, *76* (18), 5491–5500.

(247) Cho, H.; Du, X.; Rizzi, J. P.; Liberzon, E.; Chakraborty, A. A.; Gao, W.; Carvo, I.; Signoretti, S.; Bruick, R.; Josey, J. A.; Wallace, E. M.; Kaelin Jr, W. G. On-Target Efficacy of a HIF2α Antagonist in Preclinical Kidney Cancer Models. *Nature* 2016, *539*, 107–111.

(248) Corrêa, F.; Key, J.; Kuhlman, B.; Gardner, K. H. Computational Repacking of HIF-2α Cavity Replaces Water-Based Stabilized Core. *Structure* 2016, *24* (11), 1918–1927.

(249) Jacobson, K. A. New Paradigms in GPCR Drug Discovery. *Biochem. Pharmacol.* 2015, *98* (4), 541–555.

(250) Wu, B.; Chien, E. Y. T.; Mol, C. D.; Fenalti, G.; Liu, W.; Katritch, V.; Abagyan, R.; Brooun, A.; Wells, P.; Bi, F. C.; Hamel, D. J.; Kuhn, P.; Handel, T. M.; Cherezov, V.; Stevens, R. C. Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. *Science (80-.).* 2010, *330*, 1066–1071.

(251) Latorraca, N. R.; Venkatakrishnan, A. J.; Dror, R. O. GPCR Dynamics: Structures in Motion. *Chem. Rev.* 2017, *117* (1), 139–155.

(252) Lane, J. R.; May, L. T.; Parton, R. G.; Sexton, P. M.; Christopoulos, A. A Kinetic View of GPCR Allostery and Biased Agonism. *Nat. Chem. Biol.* 2017, *13*, 929–937.

(253) Bouvier, M. Oligomerization of G-Protein-Coupled Transmitter Receptors. *Nat. Rev. Neurosci.* 2001, *2* (4), 274–286.

(254) Milligan, G.; Bouvier, M. Methods to Monitor the Quaternary Structure of G Protein-Coupled Receptors. *FEBS J.* 2005, *272* (12), 2914–2925.

(255) Ferrè, S.; Baler, R.; Bouvier, M.; Caron, M. G.; Devi, L. A.; Durroux, T.; Fuxe, K.; George, S. R.; Javitch, J. A.; Lohse, M. J.; Mackie, K.; Milligan, G.; Pfleger, K. D.; Pin, J.-P. Building a New Conceptual Framework for Receptor Heteromers Sergi. *Nat. Chem. Biol.* 2009, *5*, 131–134.

(256) Ferre, S.; Casado, V.; Devi, L. A.; Filizola, M.; Jockers, R.; Lohse, M. J.; Milligan, G.; Pin, J.-P.; Guitart, X. G Protein-Coupled Receptor

Oligomerization Revisited: Functional and Pharmacological Perspectives. *Pharmacol. Rev.* 2014, *66* (2), 413–434.

(257) Wu, H.; Wacker, D.; Mileni, M.; Katritch, V.; Han, G. W.; Vardy, E.; Liu, W.; Thompson, A. A.; Huang, X.-P.; Carroll, F. I.; Mascarella, S. W.; Westkaemper, R. B.; Mosier, P. D.; Roth, B. L.; Cherezov, V.; Stevens, R. C. Structure of the Human κ-Opioid Receptor in Complex with JDTic. *Nature* 2012, *485* (7398), 327–332.

(258) Manglik, A.; Kruse, A. C.; Kobilka, T. S.; Thian, F. S.; Mathiesen, J. M.; Sunahara, R. K.; Pardo, L.; Weis, W. I.; Kobilka, B. K.; Granier, S. Crystal Structure of the M-Opioid Receptor Bound to a Morphinan Antagonist. *Nature* 2012, *485* (7398), 321–326.

(259) Huang, J.; Chen, S.; Zhang, J. J.; Huang, X.-Y. Crystal Structure of Oligomeric β1-Adrenergic G Protein–coupled Receptors in Ligand-Free Basal State. *Nat. Struct. Mol. Biol.* 2013, *20* (4), 419–425.

(260) Wang, C.; Wu, H.; Katritch, V.; Han, G. W.; Huang, X.-P.; Liu, W.; Siu, F. Y.; Roth, B. L.; Cherezov, V.; Stevens, R. C. Structure of the Human Smoothened Receptor Bound to an Antitumour Agent. *Nature* 2013, *497* (7449), 338–343.

(261) Hebert, T. E.; Moffett, S.; Morello, J. P.; Loisel, T. P.; Bichet, D. G.; Barret, C.; Bouvier, M. A Peptide Derived from a beta2-Adrenergic Receptor Transmembrane Domain Inhibits Both Receptor Dimerization and Activation. *J. Biol. Chem.* 1996, *271* (27), 16384–16392.

(262) Banères, J. L.; Parello, J. Structure-Based Analysis of GPCR Function: Evidence for a Novel Pentameric Assembly between the Dimeric Leukotriene B4 Receptor BLT1 and the G-Protein. *J. Mol. Biol.* 2003, *329* (4), 815–829.

(263) Guo, W.; Shi, L.; Filizola, M.; Weinstein, H.; Javitch, J. A. From The Cover: Crosstalk in G Protein-Coupled Receptors: Changes at the Transmembrane Homodimer Interface Determine Activation. *Proc. Natl. Acad. Sci.* 2005, *102* (48), 17495–17500.

(264) Mancia, F.; Assur, Z.; Herman, A. G.; Siegel, R.; Hendrickson, W. A. Ligand Sensitivity in Dimeric Associations of the Serotonin 5HT2c Receptor. *EMBO Rep.* 2008, *9* (4), 363–369.

(265) Hu, J.; Thor, D.; Zhou, Y.; Liu, T.; Wang, Y.; McMillin, S. M.; Mistry, R.; Challiss, R. a J.; Costanzi, S.; Wess, J. Structural Aspects of M_3 Muscarinic Acetylcholine Receptor Dimer Formation and Activation. *FASEB J.* 2012, *26* (2), 604–616.

(266) Ruprecht, J. J.; Mielke, T.; Vogel, R.; Villa, C.; Schertler, G. F. Electron Crystallography Reveals the Structure of Metarhodopsin I. *EMBO J.*

2004, *23* (18), 3609–3620.

(267) Salom, D.; Lodowski, D. T.; Stenkamp, R. E.; Trong, I. L.; Golczak, M.; Jastrzebska, B.; Harris, T.; Ballesteros, J. A.; Palczewski, K. Crystal Structure of a Photoactivated Deprotonated Intermediate of Rhodopsin. *Proc. Natl. Acad. Sci.* 2006, *103* (44), 16123–16128.

(268) Park, J. H.; Scheerer, P.; Hofmann, K. P.; Choe, H.-W.; Ernst, O. P. Crystal Structure of the Ligand-Free G-Protein-Coupled Receptor Opsin. *Nature* 2008, *454* (7201), 183–187.

(269) May, L. T.; Bridge, L. J.; Stoddart, L. A.; Briddon, S. J.; Hill, S. J. Allosteric Interactions across Native Adenosine-A3 Receptor Homodimers: Quantification Using Single-Cell Ligand-Binding Kinetics. *FASEB J.* 2011, *25* (10), 3465–3476.

(270) Brea, J.; Castro, M.; Giraldo, J.; Lopez-Gimenez, J. F.; Padin, J. F.; Cadavid, M. I.; Vilaro, M. T.; Mengod, G.; Berg, K. a.; Clarke, W. P.; Vilardaga, J.; Milligan, G.; Loza, M. I. Evidence for Distinct Antagonist-Revealed Functional States of 5HT2A Homodimers. *Mol. Pharmacol.* 2009, *75* (6), 1380–1391.

(271) Albizu, L.; Cottet, M.; Kralikova, M.; Stoev, S.; Seyer, R.; Brabet, I.; Roux, T.; Bazin, H.; Bourrier, E.; Lamarque, L.; Breton, C.; Rives, M.-L.; Newman, A.; Javitch, J.; Trinquet, E.; Manning, M.; Pin, J.-P.; Mouillac, B.; Durroux, T. Time-Resolved FRET between GPCR Ligands Reveals Oligomers in Native Tissues. *Nat. Chem. Biol.* 2010, *6* (8), 587–594.

(272) Agnati, L. F.; Fuxe, K.; Ferrè, S. How Receptor Mosaics Decode Transmitter Signals. Possible Relevance of Cooperativity. *Trends Biochem. Sci.* 2005, *30* (4), 188–193.

(273) Periole, X. Interplay of G Protein-Coupled Receptors with the Membrane: Insights from Supra-Atomic Coarse Grain Molecular Dynamics Simulations. *Chem. Rev.* 2016, *117*, 156–185.

(274) Guixà-González, R.; Javanainen, M.; Gómez-Soler, M.; Cordobilla, B.; Domingo, J. C.; Sanz, F.; Pastor, M.; Ciruela, F.; Martinez-Seara, H.; Selent, J. Membrane Omega-3 Fatty Acids Modulate the Oligomerisation Kinetics of Adenosine A2A and Dopamine D2 Receptors. *Sci. Rep.* 2016, *6* (January), 19839.

(275) Wassenaar, T. A.; Pluhackova, K.; Moussatova, A.; Sengupta, D.; Marrink, S. J.; Tieleman, D. P.; Bockmann, R. A. High-Throughput Simulations of Dimer and Trimer Assembly of Membrane Proteins. The DAFT Approach. *J. Chem. Theory Comput.* 2015, *11*, 2278–2291.

(276) Ghosh, A.; Sonavane, U.; Joshi, R. Multiscale Modelling to Understand

the Self-Assembly Mechanism of Human β2-Adrenergic Receptor in Lipid Bilayer. *Comput. Biol. Chem.* 2014, *48*, 29–39.

(277) Prasanna, X.; Chattopadhyay, A.; Sengupta, D. Cholesterol Modulates the Dimer Interface of the β2- Adrenergic Receptor via Cholesterol Occupancy Sites. *Biophys. J.* 2014, *106*, 1290–1300.

(278) Periole, X.; Knepp, A. M.; Sakmar, T. P.; Marrink, S. J.; Huber, T. Structural Determinants of the Supramolecular Organization of G Protein-Coupled Receptors in Bilayers. *J. Am. Chem. Soc.* 2012, *134*, 10959–10965.

(279) Periole, X.; Huber, T.; Marrink, S.; Sakmar, T. P. G Protein-Coupled Receptors Self-Assemble in Dynamics Simulations of Model Bilayers. *J Am Chem Soc* 2007, *129*, 10126–10132.

(280) Johnston, J. M.; Aburi, M.; Provasi, D.; Bortolato, A.; Urizar, E.; Lambert, N. A.; Javitch, J. A.; Filizola, M. Making Structural Sense of Dimerization Interfaces of Delta Opioid Receptor Homodimers. *Biochemistry* 2011, *50* (10), 1682–1690.

(281) Provasi, D.; Johnston, J. M.; Filizola, M. Lessons from Free Energy Simulations of δ-Opioid Receptor Homodimers Involving the Fourth Transmembrane Helix. *Biochemistry* 2010, *49*, 6771–6776.

(282) Johnston, J. M.; Wang, H.; Provasi, D.; Filizola, M. Assessing the Relative Stability of Dimer Interfaces in G Protein-Coupled Receptors. *PLoS Comput. Biol.* 2012, *8*, e1002649.

(283) Lelimousin, M.; Limongelli, V.; Sansom, M. S. P. Conformational Changes in the Epidermal Growth Factor Receptor: Role of the Transmembrane Domain Investigated by Coarse-Grained MetaDynamics Free Energy Calculations. *J. Am. Chem. Soc.* 2016, *138* (33), 10611–10622.

(284) Livingston, M.; Heaney, L. G.; Ennis, M. Adenosine , Inflammation and Asthma – a Review. *Inflamm. Res.* 2004, *53*, 171–178.

(285) Goldberg, S. R.; Mallol, J.; Lopez-gimenez, J. F.; Ciruela, F.; Casado, V.; Rodrigues, R. J.; Luja, R.; Burguen, J.; Borycz, J.; Rebola, N.; Goldberg, S. R.; Mallol, J.; Corte, A.; Canela, E. I.; Lo, J. F.; Milligan, G.; Lluis, C.; Cunha, R. A.; Ferre, S.; Franco, R. Presynaptic Control of Striatal Glutamatergic Neurotransmission by Adenosine A1-A2A Receptor Heteromers. *J. Neurosci.* 2006, *26* (7), 2080–2087.

(286) Ferre, S.; Bonaventura, J.; Tomasi, D.; Navarro, G.; Moreno, E.; Cortés, A.; Lluís, C.; Casadó, V.; Volkow, N. D. Allosteric Mechanisms within the Adenosine A 2A E Dopamine D 2 Receptor Heterotetramer. *Neuropharmacology* 2016, *104*, 154–160.

(287) Torvinen, M.; Marcellino, D.; Canals, M.; Agnati, L. F.; Lluis, C.; Franco, R.; Fuxe, K. Adenosine A 2A Receptor and Dopamine D 3 Receptor Interactions : Evidence of Functional A 2A / D 3 Heteromeric Complexes. *Mol. Pharmacol.* 2005, *67*, 400–407.

(288) Ferré, S.; Karcz-Kubicha, M.; Hope, B. T.; Popoli, P.; Burgueño, J.; Gutiérrez, M. A.; Casadó, V.; Fuxe, K.; Goldberg, S. R.; Lluis, C.; Franco, R.; Ciruela, F. Synergistic Interaction between Adenosine A2A and Glutamate mGlu5 Receptors: Implications for Striatal Neuronal Function. *Proc. Natl. Acad. Sci. U. S. A.* 2002, *99* (18), 11940–11945.

(289) Ferré, S.; Goldberg, S. R.; Lluis, C.; Franco, R. Looking for the Role of Cannabinoid Receptor Heteromers in Striatal Function. *Neuropharmacology* 2009, *56* (SUPPL. 1), 226–234.

(290) Marcellino, D.; Carriba, P.; Filip, M.; Borgkvist, A.; Frankowska, M.; Bellido, I.; Tanganelli, S.; Mu, C. E.; Fisone, G.; Lluis, C.; Agnati, L. F.; Franco, R.; Fuxe, K. Antagonistic Cannabinoid CB 1 / Dopamine D 2 Receptor Interactions in Striatal CB 1 / D 2 Heteromers . A Combined Neurochemical and Behavioral Analysis. *Neuropharmacology* 2008, *54*, 815–823.

(291) Canals, M.; Burgueño, J.; Marcellino, D.; Cabello, N.; Canela, E. I.; Mallol, J.; Agnati, L.; Ferré, S.; Bouvier, M.; Fuxe, K.; Ciruela, F.; Lluis, C.; Franco, R. Homodimerization of Adenosine A2A Receptors: Qualitative and Quantitative Assessment by Fluorescence and Bioluminescence Energy Transfer. *J. Neurochem.* 2004, *88* (3), 726–734.

(292) Lebon, G.; Warne, T.; Edwards, P. C.; Bennett, K.; Langmead, C. J.; Leslie, A. G. W.; Tate, C. G. Agonist-Bound Adenosine A2A Receptor Structures Reveal Common Features of GPCR Activation. *Nat. Lett.* 2011, *474*, 521–525.

(293) Sali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* 1993, *234*, 779–815.

(294) Robertson, N.; Errey, J. C.; Ng, I.; Hollenstein, K.; Tehan, B.; Hurrell, E.; Dore, A. S.; Bennett, K.; Congreve, M.; Magnani, F.; Tate, C. G.; Weir, M.; Marshall, F. H. Article Structure of the Adenosine A 2A Receptor in Complex with ZM241385 and the Xanthines XAC and Caffeine. *Structure* 2011, *19*, 1283–1293.

(295) Wassenaar, T. A.; Ingólfsson, H. I.; Böckmann, R. A.; Tieleman, D. P.; Marrink, S. J. Computational Lipidomics with Insane: A Versatile Tool for Generating Custom Membranes for Molecular Simulations. *J. Chem. Theory Comput.* 2015, *11* (5), 2144–2155.

(296) Wassenaar, T. A.; Pluhackova, K.; Böckmann, R. A.; Marrink, S. J.; Tieleman, D. P. Going Backward: A Flexible Geometric Approach to

Reverse Transformation from Coarse Grained to Atomistic Models. *J. Chem. Theory Comput.* 2014, *10* (2), 676–690.

(297) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* 2004, *25* (13), 1605–1612.

(298) Choe, H.-W.; Kim, Y. J.; Park, J. H.; Morizumi, T.; Pai, E. F.; Krauß, N.; Hofmann, K. P.; Scheerer, P.; Ernst, O. P. Crystal Structure of Metarhodopsin II. *Nature* 2011, *471* (7340), 651–655.

(299) Schertler, G. F.; Hargrave, P. A.; Hargravet, P. A. Projection Structure of Frog Rhodopsin in Two Crystal Forms. *Proc Natl Acad Sci U S A* 1995, *92* (25), 11578–11582.

(300) Filizola, M. Increasingly Accurate Dynamic Molecular Models of G-Protein Coupled Receptor Oligomers: Panacea or Pandora's Box for Novel Drug Discovery? *Life Sci.* 2010, *86* (15–16), 590–597.

(301) Guo, W.; Urizar, E.; Kralikova, M.; Mobarec, J. C.; Shi, L.; Filizola, M.; Javitch, J. A. Dopamine D2 Receptors Form Higher Order Oligomers at Physiological Expression Levels. *EMBO J.* 2008, *27* (17), 2293–2304.

(302) González-Maeso, J.; Ang, R. L.; Yuen, T.; Chan, P.; Weisstaub, N. V.; López-Giménez, J. F.; Zhou, M.; Okawa, Y.; Callado, L. F.; Milligan, G.; Gingrich, J. A.; Filizola, M.; Meana, J. J.; Sealfon, S. C. Identification of a Serotonin/glutamate Receptor Complex Implicated in Psychosis. *Nature* 2008, *452* (7183), 93–97.

(303) Ye, L.; Van Eps, N.; Zimmer, M.; Ernst, O. P.; Scott Prosser, R. Activation of the A2A Adenosine G-Protein-Coupled Receptor by Conformational Selection. *Nature* 2016, *533* (7602), 265–268.

"It is our choices, Harry, that show what we truly are,
far more than our abilities."

Albus Dumbledore – JK Rowling
Harry Potter and The Chamber of Secrets

ACKNOWLEDGEMENT

We are, at last, at the acknowledgement section. Someone might think that this section is kind of mandatory, so taken for granted to be considered a cliché. I am instead writing these acknowledgements with sincere gratitude to all the people that have contributed to my work, and I want you to read what follows in this light.

My first thank goes to the person who has lead me through my academic journey. I should say Prof. Bonati, but I prefer to say Laura. Laura has followed me during these three years always trying to bring out the best from me and getting me back on track when necessary. She supervised me putting on the foreground my personal growth, building a special relationship with all the components of the group. Thanks for your way of tutoring me in this years.

My second thanks go to my family. They have always supported me and encouraged me in my study. They were always present when I needed, and ready to help me without I needed to ask. I only could reach my goal thanks to the peacefulness that they provided me. Thank you mum, thank you dad. I could not ask more.

Talking about family I must mention Raffaella. My partner, my future wife. She went through the same path I am traveling and knows what this mean. Thanks for being always present and bearing my upset moments, thanks for the advice, the laughs, and all the evening spent together.

My PhD work could not be possible without my lab mates. I should thank Sara, for grateful discussion and for sharing her Biology knowledge with me. She has strongly contributed to my personal growth. A thank goes to Dario and Domenico, with which I shared a

part of my PhD and that preceding me have contributes to the growth of the lab. I would also thanks Lara, for all her mother food, and for giving me the opportunity of teaching to someone eager of learn.

During these three years I was so lucky to have had two experiences abroad. I worked with Dr. Alessandro Pandini, which played the role of *Ze Boss* but with the affection of an older brother. Thanks for all the advice and the attention that you gave me.

An important part of my PhD work was spent at USI at Lugano, where I met fantastic people. Prof. Limongelli has built a stimulating environment and offered me the opportunity to work side by side with Daniele, Stefano and Simone. Thank you all for have welcomed me into your group.

My sincere thanks also go to all the people that has spent their time for the good success of my work. My co-tutors Prof. G. Colombo and L. De Gioia, for fruitful discussion during final years exams; the external reviewers of my thesis Dr. A. Fornili and Prof. C. Camilloni for the careful review, and Dr. C. Minici for the collaboration on Chapter 4 work.

A great thanks to all the people that supported me outside the work. Thanks to Raffaella's family that welcomed me within their world, to all my friends, in particular Valerio, Diego, Clelia, Federico, Cristian, Antonella, Maria, Mattia, Leo, Alessio, Daniela, Lara and Francesco for all the time spent together.

Last but not least I would like to thank a person that has silently supported with all the church candle lit up for me, all her prayers and all goodies I've always found in her house thanks granny.