THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Discovering and Exploiting Hidden Pockets at Protein Interfaces

Thesis for the Degree of Doctor of Philosophy

Rémi Cuchillo

THE UNIVERSITY
of EDINBURGH

School of Chemistry

First Supervisor : Dr Julien Michel
Second Supervisor : Dr Philip Camp

Submitted: 01/12/2014

Assessment committee:

Internal Examinator:

Dr. Carole A Morrison

External Examinator:

Dr. Xavier Barril

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

Rémi Cuchillo

Abstract

The number of three-dimensional structures of potential protein targets available in several platforms such as the Protein Data Bank is subjected to a constant increase over the last decades. This observation should be an additional motivation to use structure-based methodologies in drug discovery. In the recent years, different success stories of Structure Based Drug Design approach have been reported. However, it has also been shown that a lack of druggability is one of the major causes of failure in the development of a new compound.The concept of druggability can be used to describe proteins with the capability to bind drug-like compounds. A general consensus suggests that around 10% of the human genome codes for molecular targets that can be considered as druggable.

Over the years, the protein druggability was studied with a particular interest to capture structural descriptors in order to develop computational methodologies for druggability assessment. Different computational methods have been published to detect and evaluate potential binding sites at protein surfaces. The majority of methods currently available are designed to assess druggability of a static structure. However it is well known that sometimes a few local rearrangements around the binding site can profoundly influence the affinity of a small molecule to its target. The use of techniques such as molecular dynamics (MD) or Metadynamics could be an interesting way to simulate those variations.

The goal of this thesis was to design a new computational approach, called JEDI, for druggability assessment using a combination of empirical descriptors that can be collected 'on-the-fly' during MD simulations. JEDI is a grid-based approach able to perform the druggability assessment of a binding site in only a few seconds making it one of the fastest methodologies in the field. Agreement between computed and experimental druggability estimates is comparable to literature alternatives. In addition, the estimator is less sensitive than existing methodologies to small structural rearrangements and gives consistent druggability predictions for similar structures of the same protein. Since the JEDI function is continuous and differentiable, the druggability potential can be used as collective

variable to rapidly detect cryptic druggable binding sites in proteins with a variety of MD free energy methods.

Acknowledgements

It is a pleasure for me to express my gratitude to the many people who have helped me or have been an important part of my life the last three years.

First of all, I am grateful to my supervisor, Dr. Julien Michel, for his patience, motivation, enthusiasm, friendly advice and constructive criticism. His guidance helped me in all the time of research and writing of this thesis. I would like to thank him for entrusting me with this very exciting challenge.

I am also thankful to my second supervisor Dr. Philip Camp, and all members of the computational chemistry office. I would like to thanks previous and current members of the Michel research group, and in particular George Gerogiokas, Haris Georgiou, Juan Bueren Calabuig. Very special thanks go to Gaetano Calabro. I really enjoyed to work in front you ;-) I am very happy for the friendship we have build up during the years. Tanti auguri per la stesura della tesi!

I guess it is time to thank my friends from 'Edinbra'. I would like to start with Kevin Pinto-Gil. I really enjoyed the time we spent together. He tenido mucha suerte de conocerte en los cursos del IELTS. Que la fuerza te acompañe ! Je tiens à remercier mon colocataire Yann Rimbaud. Je pense que ma dernière année fut la meilleure principalement grâce à (ou à cause de) toi. Un grand merci pour tous ces bons moments. Je te souhaite bon courage pour le reste de ta thèse. J'ai presque failli oublier Irène Cadavid ... J'aurai pu également t'inclure comme colocataire mais le 'council' ne considère pas le canapé comme une chambre... On aura quand même bien rigolé pendant trois ans, que des bons souvenirs.

I start a new paragraph to thank my friends from Paris. Je voudrais commencer par mes deux amis de toujours Mamadou Gassama et Radwan Zerbib qui malgré la distance ont toujours été présent. Vous savoir au chaud sous le soleil africain m'a beaucoup aidé pendant les hivers écossais. Un grand merci à Rawane, David, Rémi, ... et tous les autres pour m'avoir changé les idées à chacun de mes retours sur Paris. Pour terminer, je voudrais remercier tout particulièrement Harry Primack pour sa simplicité (dans le bon sens du terme) et pour m'avoir supporté surtout ces derniers temps.

To conclude this section, I would like to thanks my familly. Un grand merci à mes parents et à ma soeur pour leur soutien sans faille. Je sais que ca n'a pas toujours été facile pour vous ces dernières années mais c'est grâce à vous si j'en suis là aujourd'hui. Merci pour tout !

Por fin, quiero agradecer Valentina por su gentileza y apoyo. Sé que no todo ha sido facil especialmente durante la redaccion de la tesis. Tu presencia fue una gran ayuda.

Rémi Cuchillo

Edinburgh, Scotland, December 2014

Contents

List of Abbreviations

A . . . . . Alanine

ADMETox . Absorption Distribution Metabolism Elimination Toxicology

Apo . . . . . Without ligand

BEMD . . . . Bias-Exchange Metadynamics

bHLHZip . . basic Helix Loop Helix leucine Zipper

C . . . . . Cysteine

CD . . . . . Circular Dichroism

CV . . . . . Collective Variable

D . . . . . Aspartic acid

DCD . . . . . Druggable Cavity Directory

E . . . . . Glutamic acid

F . . . . . Phenylalanine

FEP . . . . . Free Energy Profile

G . . . . . Glycine

H . . . . . Histidine

HIV . . . . . Human Immunodeficiency Virus

Holo . . . . . With ligand

hPNMT . . . human Phenylethanolamine N-Methyltransferase

HTS . . . . . High-Throughput Screening

I . . . . . Isoleucine

IDP . . . . . Intrinsically Disordered Protein

JEDI . . . . . Just Exploring Druggability at protein Interfaces

K . . . . . Lysine

L . . . . . Leucine

M . . . . . Methionine

MD . . . . . Molecular Dynamics

N . . . . . Asparagine

NMR . . . . . Nuclear Magnetic Resonance

NRDD . . . . Non Redundant Druggability Dataset

P . . . . . Proline

PDB . . . . . Protein Data Bank

Q . . . . . Glutamine

R . . . . . Arginine

RMSD . . . . Root-Mean-Square Deviation

S . . . . . Serine

SBDD . . . . Structure-Based Drug Design

SpeB . . . . . Streptococcal pyrogenic exotoxin B

T . . . . . Threonine

Tau . . . . . Tubule-associated unit

US . . . . . Umbrella Sampling

V . . . . . Valine

VHL . . . . . Von Hippel-Lindau

W . . . . . Triptophan

WHAM . . . Weighted Histogram Analysis Method

Y . . . . . Tyrosine

3D . . . . . Three Dimensional

1

Introduction

The first chapter introduces the motivation, the theoretical, experimental and computational concepts used throughout the thesis in the context of target-based drug discovery.

In the past, the drug discovery process was mainly based on screening of natural products based on the success of traditional medicines and herbal remedies.[1] Over the last 30 years, only one third of new drugs approved annually by the Food and Drug Administration came from natural products with sometimes semi-synthetic modifications.[2, 3] Therefore, the discovery of new structures derived from natural products that have not been already registered, is becoming more difficult, encouraging the pharmaceutical industry to develop innovative strategies.[4–6] Progress achieved in the field of chemical synthesis and pharmacology have led to a systematic approach to develop rapidly new drug candidates with a greater efficiency.[7] During the successive phases of the modern drug discovery process, thousands of molecules are screened and tested for desirable properties in preclinical models of diseases, leading to a very small number of drug candidates tested in clinical trials (Figure 1.1). The elapsed time between the identification of a relevant biological target and the marketing of a new drug is about 14 years with total research and development costs superior to $ 800 million.[7–9]

Therapeutic targets are usually single proteins or complexes involved in disease. A rough estimation suggests that all the commercialized drugs to date target between 300 and 500 different proteins.[10–12] However, these protein drug targets are not equally distributed among the proteome.[13] Certain families of proteins are more represented in the human genome, or more frequently involved in pathological pathways. However it may be difficult to identify a drug that target a protein without undesirable side-effects if there exist a large number of homologous proteins. In addition, some proteins are simply easier to target than others. For instance, it is easier for a pharmaceutical compound to perturb the interactions between a small molecule and a protein binding site than to disrupt protein-protein interactions.[14, 15] In the past, the identification of a therapeutic target relied on empirical evidences.[16] Since significant advances have been made

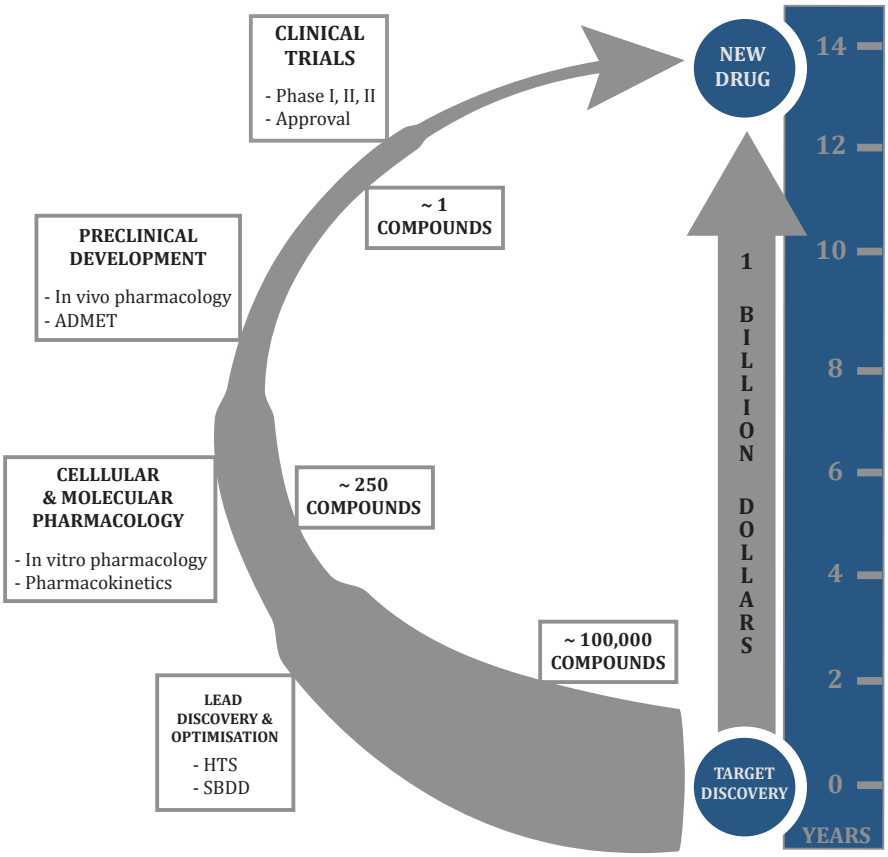

Figure 1.1: Target-based drug discovery process. This process takes place over a period of 12-15 years and represents an investment around $ 800 million. The approval of a new medication is usually the result of different experimental procedures that involve the testing of thousands of compounds. Based on several criteria such as the affinity and the selectivity for a protein target but also the toxicity and the efficiency, the initial set of small molecules is then drastically refined and carefully optimized. Only a few drug candidates are finally selected for further clinical trials.
HTS: High-Throughput Screening, SBDD: Structure-Based Drug Design, ADMET: Absorption Distribution Metabolism Elimination Toxicology.

in the fields of genomics, proteomics and bioinformatics, it is now possible to identify more specifically genes and proteins that are involved in diseases but also the more likely to become a therapeutic target.[17, 18]

Once a protein target has been selected, the next step of the target-based process consists in identifying diverse small molecules able to bind the protein of interest. These molecules are usually discovered using screening techniques such as High-Throughput Screening (HTS) of large libraries containing up to hundred thousands of compounds. However, because such approaches are usually time consuming and expensive, knowledge-based methodologies are sometimes preferred.[15, 19] In recent years, virtual screening strategies, or screening *in silico*, have been introduced as an alternative or complementary method to guide HTS. These techniques are generally fairly easy to use and at a much lower cost than experimental screenings.[20] In addition, the constant evolution of technology has dramatically reduced the computational cost required for the simulation of complex systems or for the query of databases of several millions of molecules. Virtual screening is now used in many projects to select, within large libraries of molecules, a limited number of compounds that will be screened experimentally accelerating the hit identification.[21, 22] Then, a number of hits are carefully selected for the lead generation and lead optimization phase. The choice of these compounds is mainly based on the chemical structure and the affinity between the ligand and its target. A good hit is usually a small molecule with an affinity between 100 nM and 5 μM, a scaffold allowing to graft several substituents and an overall chemical structure different from the pharmaceutical patents already registered.[23, 24] In the case of fragment-based drug design, a weaker binding affinity may be observed.[25] The hit-to-lead optimization step aims to increase the affinity of a compound to its target to reach a dissociation constant in the order of

the nM range. The lead optimization attempts to maintain sufficient specificity towards other proteins. Additional parameters have to be taken into account to finally propose few 'drug-like' molecules meeting the Absorption Distribution Metabolism Elimination Toxicology (ADMETox) criteria.[26–28] Owing to the complexity of these steps, they are often considered as the most critical in drug discovery.[29] After this process, different phases of pre-clinical and clinical studies are performed during which the safety and the efficacy of all drug candidates are evaluated directly from trials in animals and patients respectively.[30]

1.1 Structure-based Drug Design

Recent advances in the field of structural biology bring a new dimension to the characterization of therapeutic molecules. The Structure Based Drug Design approach (SBDD) is an iterative process exploiting the physicochemical properties extracted from a three dimensional (3D) protein structure, preferentially interacting with a ligand, to design or optimize potential drug candidates (Figure 1.2).[31]

Accordingly, the structure determination of the protein target is a crucial step in SBDD. Several approaches can be used. X-ray crystallography is currently the method of choice. Different orientations of a crystal containing a continuous arrangement of a specific protein conformation are exposed to a X-ray beam. Then, the protein structure is reconstructed from the diffraction pattern obtained by the X-ray scattering caused by the molecules inside the crystal.[32] The technique, substantially improved since the advent of structural genomics and

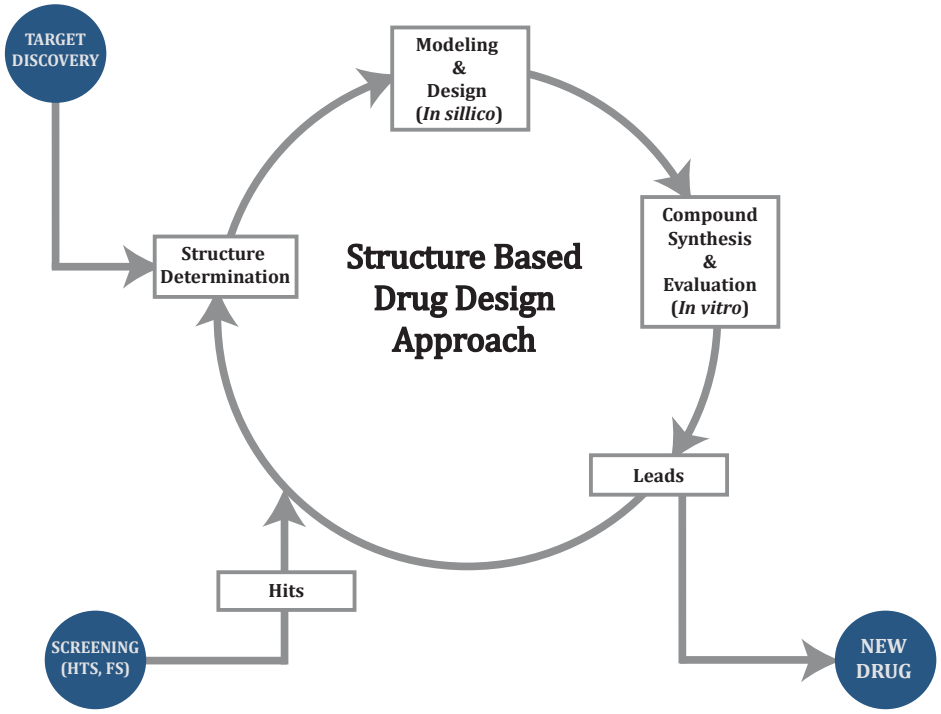

Figure 1.2: Structure-Based Drug Design approach. The SBDD approach aims to optimize several hit molecules identified using diverse techniques such as HTS. During this iterative process, the 3D structure of the protein target interacting with a ligand is used to increase specific ligand parameters such as the affinity or the selectivity between the two partners. At the end, the optimized ligands presenting the characteristics of potential drug candidates are called leads.

synchrotron radiation, has allowed solving the 3D structure of very large systems such as the eukaryotic ribosome.[33] In complement to crystallography, Nuclear Magnetic Resonance (NMR) plays an increasing role in the structural analysis of macromolecules in solution.[34] NMR is a spectroscopic technique exploiting the magnetic properties of atomic nuclei which absorb electromagnetic radiation emitted at a specific frequency in the presence of a strong magnetic field. The analysis of the observed frequency shift gives information about the environment of the considered atom and allows the reconstruction of the structure step by step. This technique does not require protein crystallization but is limited by other constraints such as the size of the molecule being studied (only proteins with low molecular weight) or its solubility. However, NMR offers the possibility to study the dynamics of molecules, the interactions between macromolecules and solvent and also structural changes that occur during the formation of transient complexes. Furthermore, the technique opens up interesting possibilities, still little exploited, for the study of macromolecules such as membrane proteins.[35] Finally, computational approaches such as homology modelling may also be used for protein structure determination.[36] The resolution of the 3D structure of the protein target is a crucial step for the success of SBDD approach. Indeed, structures based on electron density maps at 1.2 Å resolution correspond to an atomistic resolution allowing to characterize, without ambiguity, interactions between a ligand and its target such as the presence of hydrogen bonding interactions. Structures solved with a resolution higher than 3.0 Å are usually much less suitable for these detailed analyses.[5]

Given a high-resolution structure, the protein is analysed to provide an understanding of the binding mode between the ligand and its target. This information is used to increase the affinity between the two partners by modifying the chemical scaffold of the ligand or by introducing new substituents. The process

is repeated until a series of lead molecules are obtained. A good lead molecule will typically have an affinity in the nM range and a sufficient selectivity against the target while respecting ADEMtox properties required for a drug candidate.

Historically, the first successes in the SBDD approach led to the development of small molecules interacting with intracellular proteins such as protease inhibitors of the AIDS virus (HIV), as well as molecules limiting the flu virulence factor.[37, 38] Such examples have multiplied in the last ten years to include other types of more complex molecules such as antibodies targeting extracellular or exogenous proteins.[39] This approach is also used to define the immunogenic domains of viral proteins (shell of the virus) in order to develop new vaccines. This more rational methodology is today an essential step to design more effective drug candidates while also reducing both the timeline and the cost of the drug discovery process.

1.2 Druggability

The number of 3D structures of potential protein targets available in several platforms such as the Protein Data Bank (PDB) is constantly increasing.[40] This observation should be an additional motivation to use structure-based methodologies in drug discovery. Over the last decades, around 60% of drug discovery projects failed to identify viable leads able to modulate the activity of a protein target due to a lack of druggability.[4, 41, 42]

1.2.1 Definition

Druggability has been used in a large number of publications in different fields to describe in different contexts the properties of genes, ligands or proteins. Thus, the term is sometimes ambiguous.[43] In this thesis, druggability is applied to a protein target. Analyses of the sequenced human genome indicate that less than 50% of disease-involved genes code for druggable proteins.[44, 45] When assessing protein druggability in target validation, one is often focused on the capability of a therapeutic target to bind a drug-like small molecule, leaving aside many important facets of the drug discovery and development process such as selectivity.[44] Therefore, protein druggability is closely related to the definition of drug-likeness in this context. Historically, a drug-like compound is a molecule meeting at least three of the criteria laid down by Lipinski's Rule of Five:[27, 46]

- no more than 5 hydrogen bond donors.
- no more than 10 hydrogen bond acceptors.
- A molecular mass less than 500 daltons.
- An octanol-water partition coefficient (log P) not greater than 5.

Over the years, the characteristics of compounds presenting a good oral bioavailability was refined, and other parameters such as the number of rotatable bonds or aromatic rings were also found to play a significant role in the druglikeness.[47, 48] However, those rules should not be considered as well established but more as a guideline.[49] The same shall apply to the definition of protein binding site druggability. A druggable cavity tends usually to be a buried pocket, more hydrophobic than hydrophilic and large enough to bind a small molecule able to

modulate the protein activity. The definition of a nondruggable protein is also questionable. Indeed, this term is not only applied to protein binding sites that do not respect the druggability guidelines. The nondruggable target definition covers also proteins binding a drug-like molecule with a high affinity but not able to induce a therapeutic effect despite intensive efforts. In addition to druggable and nondruggable proteins, a new category of druggable protein targets called 'difficult' has recently been introduced.[50] It was suggested that this category of proteins should be targeted with highly polar molecules administrated as pro-drugs. Since druggability is closely linked to the notion of binding site in this specific context, the terms 'bindability' or 'ligandability' have been recently introduced to avoid ambiguities.[51, 52] In the rest of this thesis, the term druggability is used to describe the capability of protein target to bind a drug-like compound.

1.2.2 Existing Methodologies

With a growing interest in evaluating the capacity of a protein target to bind a drug-like compound with a high affinity, several studies have focused on developing computational methodologies that correlate structural descriptors to this property. An early effort was contributed by Hajduk and coworkers.[53] NMR-based fragment screening was used to develop a mathematical model for druggability measurements using empirical descriptors correlated to NMR hit rates. The methodology relies on the assumption that a druggable cavity tends to bind more fragments than a nondruggable pocket. Based on the insight II software to detect protein binding pockets, six structural descriptors (surface area, polar & apolar contact area, the third & first principal component capturing

the shape of the cavity and the pocket compactness) were found to correlate with NMR hit rates. A second approach, called MAP_{POD} (Equation 1.1), was published by Cheng *et al.* shortly after.[50] The authors proposed a scoring function to assess the maximal affinity between a small molecule and a binding site based on physicochemical and geometric features.

$$\Delta G_{MAP_{POD}} \approx -\gamma(r) A^{target}_{nonpolar} \frac{A^{target}_{druglike}}{A^{target}_{total}} + C \tag{1.1}$$

where $\gamma(r)$ describes the curvature of the binding site, $A^{target}_{nonpolar}$ & A^{target}_{total} captures the apolar surface area and the total surface area respectively, $A^{target}_{druglike}$ is fixed to 300 $Å^2$ and C is a constant. The model was derived from the first publicly available protein dataset compiled for the purpose of druggability studies. This small dataset gathers 63 crystallographic structures of 27 different proteins that have been subjected to past structure-based drug design campaigns. Protein targets interacting with a commercialized drug were considered as druggable whereas proteins without any drug on the market despite intensive efforts to develop a medication are considered as non druggable. These approaches have paved the way for the development of different computational methods that aim to detect and evaluate potential binding sites at protein surfaces.

The public dataset compiled for MAP_{POD} was used to parameterize D_{score} (Equation 1.2), a druggability function coupled with the pocket detector SiteMap.[54, 55] D_{score} is a simple linear combination of three descriptors reflecting the volume, enclosure and hydrophobicity of the binding site.

$$D_{score} = 0.094\sqrt{n} + 0.6e - 0.324p \tag{1.2}$$

where n is the number of site points, e is the degree of enclosure and p captures

the hydrophobicity.
This approach was found to discriminate the three categories of protein targets introduced by Cheng *et al.* One of the main limitations of D_{score} is the execution time. The method relies on expensive grid point energy calculations that may significantly slow down the druggability predictions. Therefore, D_{score} might not be suitable for high throughput aplication.

To overcome this limitation, the fpocket has been developed.[56, 57] This methodology is able to assess protein druggability on very large dataset at a reasonable computational cost, and is essantially based on hydrophobicity and polarity predictions (Equation 1.3).

$$drugscore(z) = \frac{e^{-z}}{1 + e^{-z}} \tag{1.3}$$

where z is a linear combination of three descriptors: the normalized mean local hydrophobic density, the pocket hydrophobicity score and the normalized polarity score.
fpocket was trained and validated on a large dataset of 70 unique proteins publicly available. In addition to distinguish druggable, difficult and nondruggable proteins, the approach is one of the fastest in the field providing an interesting tool for virtual screening. However, the pocket druggability predictions at the protein surface are dependent on each other. Indeed, the mean local hydrophobic density is normalized compared to other binding sites on the same protein. Consequently, fpocket is not tailored to perform post-processing druggability assessment from structures obtained using computational tools capturing protein flexibility in solution such a molecular dynamics (MD) simulation.

More recently, MD-based methodologies have been introduced.[58–60] One of the first methods based on first-principles molecular simulations was published

by Seco and coworkers.[58] In this grid-based approach, an explicit restrained MD simulation of a protein is performed in the presence of a given concentration of isopropyl alcohol. The binding propensities of the probe at the protein surface are then back-computed to evaluate a binding free energy (Equation 1.4).

$$\Delta G_i = -k_B T ln \frac{N_i}{N_0} \tag{1.4}$$

where k_B is the Boltzman constant, T is the temperature, N_i is the observed population and N_0 is the expected population.

A similar protocol was recently applied on different systems using several kinds of probes without any restraints on the protein.[59] The authors showed that probe molecules could induce both local and global structural rearrangements increasing the target druggability. However, all these techniques can generate a large number of false positives or denature the protein at high probe concentrations, requiring the judicious use of positional restraints to limit the occurrence of undesirable conformational changes. Also, probe diffusion necessary to compute occupancies to buried cavities can be very slow with standard MD approaches. To overcome the limitations described previously, this thesis introduces a new grid-based methodology, called JEDI (Just Exploring Druggability at protein Interfaces), to assess protein druggability during a MD simulation. The entire process is described in details in the chapters 4 & 5.

1.2.3 Protein Flexibility and Druggability

Protein binding site flexibility has been found to play an important role in the binding process with a small molecule.[61] This flexibility may involve small or

large structural rearrangements. For instance, motions of few amino acids located near the active site of acetylcholinesterase have been reported.[62] Because of its involvement in the memorization process and Alzheimer's disease, this enzyme has been extensively studied.[63, 64] Using MD simulations, the authors were able to identify two residues of the active site showing a high flexibility. Some proteins show larger structural rearrangements to adopt an active form such as streptococcal pyrogenic exotoxin B (SpeB). SpeB is a cysteine protease that is secreted as an inactive zymogen (precursor protein of an enzyme).[65] As with many proteases, the activation of SpeB involves a proteolytic digestion that releases a pro-domain to form the active enzyme. The displacement of the pro-domain induces a large intramolecular rearrangement. A loop moves over 25 Å from one pole to the opposite pole of the protein. A second loop, which contains the catalytic histidine, is then free to move away from the substrate binding site. The active conformation of the enzyme is formed and the protein may perform its function. More recently, Alvarez-Garcia *et al* have studied the impact of protein flexibility on binding free energy using MD simulations. Results suggest that an accurate binding free energy prediction requires to consider binding flexibility. Furthermore, they highlighted that the use of soft positional restraints may be an interesting approach allowing to sample significant local structural rearrangement at a reasonable computational cost.[66]

1.3 Classical Force Fields

The usefulness of a high-resolution 3D structure for a protein has been discussed in the previous section. However, this set of cartesian coordinates

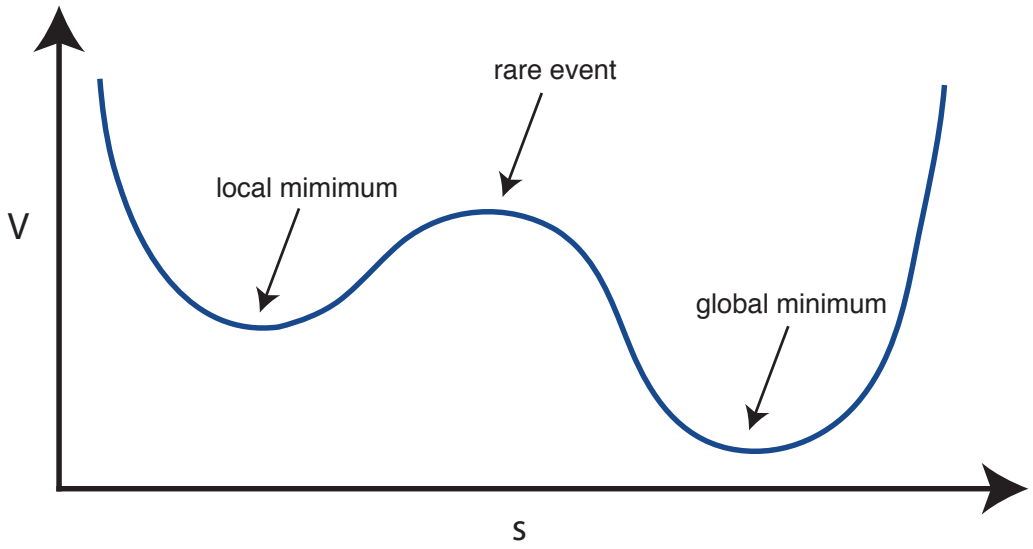

Figure 1.3: Illustation of the potential energy surface of a protein along a reaction coordinate s.

represents only one structural conformation corresponding usually to a minimum on the potential energy surface of the protein (Figure 1.2). It is well known that in solution a protein may oscillate between structurally diverse conformations of similar low energy. Nevertheless, it is often difficult to resolve with experiments all these possible alternative structures. Techniques based on MD simulations presented below represent an attractive tool to simulate those variations.

Simulations performed to explore the conformational space of a protein rely on three criteria. First, the degrees of freedom that are explicitly simulated must be defined. In this thesis, the degrees of freedom are the Cartesian coordinates of the protein atoms. Then, it is necessary to define a mathematical function to evaluate the total energy of the system for different arrangements of atoms. In classical dynamics simulations, this function is called the Hamiltonian of the system. The Hamiltonian is defined as the sum of the kinetic energy (Equation 1.5) and the potential energy (Equation 1.6).

The kinetic energy of the system is only dependent on the mass m and the velocity v of each atom.

$$K(m) = \sum_{i=1}^{n} \frac{1}{2} m_i \mathbf{v}_i^2 \tag{1.5}$$

where v_i is the velocity of atom i.

The calculation of the potential energy of a system relies on several parameters such as the mass, the charge and the distance between atoms. Those variables and the equations used to compute the potential energy are called a force field. Many different force fields are available. However, all of them are based on experimental data such as vibrational frequencies of bonds obtained by infrared spectroscopy or by measuring the bond length using X-ray crystallography. Ab initio methods also frequently provide crucial information on the twist angles or the bond vibration frequencies.

In most of force fields used in biomolecular simulations, the potential energy is described by the following equation:

$$V(\mathbf{r}) = V(\mathbf{r})^{bonded} + V(\mathbf{r})^{non-bonded} + V(\mathbf{r})^{bias} \tag{1.6}$$

where the first term corresponds to the interactions between atoms with covalent bonding including angles and dihedral angles, the second represents the van der Waals and electrostatic interactions while the last term is used in specific cases that will be described later.

The force $\mathbf{f}_i$ acting on a particle i is given by:

$$\mathbf{f}_i = -\frac{\partial V}{\partial \mathbf{r}_i} \tag{1.7}$$

With knowledge of the forces acting on the particles in the system, it is possible to solve equations of motions that predict the time evolution of the collection of particles using Lagrangian or Hamiltonian formalisms. The Newton's equations of motion have been used in this thesis

$$\mathbf{f}_i = m_i \mathbf{a}_i \tag{1.8}$$

$$\frac{d\mathbf{r}_i(t)}{dt} = \mathbf{v}_i(t) \tag{1.9}$$

$$\frac{d\mathbf{v}_i(t)}{dt} = \frac{\mathbf{f}_i(t)}{m_i} \tag{1.10}$$

where $\mathbf{v}_i$ and $\mathbf{f}_i$ are respectively the atomic velocity and the force acting on the atom i at the time t. Newton's equations are only valid for the Cartesian coordinates $\mathbf{r}_i$ of a particle with a mass m_i. The initial particle velocities are given by a Maxwell-Boltzmann distribution (Equation 1.11).

$$P(\mathbf{v}_{i,\mathbf{r}}) = \left(\frac{m_i}{2\pi k_B T}\right)^{1/2} \exp\left(-\frac{m_i \mathbf{v}_{i,\mathbf{r}}^2}{2k_B T}\right) \tag{1.11}$$

Numerical integration of the Equations 1.7, 1.9 and 1.10 are iteratively solved over the MD simulation time. The molecular system can be coupled to external variables. Indeed, an additional term $V_{bias}(\mathbf{r})$, may be used either to limit particle motions or to enhance the conformational sampling.

In the following part of the thesis, the description of the systems is based on the formalism of classical physics. By contrast with quantum physics where

electron positions around a nucleus are considered, atoms are represented as spheres with a fixed volume and a fixed partial charge.

1.3.1 The bonded interactions

Two different AMBER forcefields (AMBER99sb and AMBER99sb-ILDN) have been used in this thesis. Therefore, the description below is based on a specific formalism.

The bonded interactions are a combination of four terms:

$$V^{bonded} = V^{bond} + V^{angle} + V^{dihedral} + V^{improper} \tag{1.12}$$

They describe the bond elongations, the angle deformations and the torsions for the periodic and improper dihedral angles.

1.3.1.1 Bond-stretching term

The energy of a covalent bond between two atoms is calculated by analogy with a harmonic oscillator from the distance between two atoms (Hooke's law):

$$V^{bond} = \sum_{n=1}^{N_b} \frac{k_{b_n}}{2}(b_n - b_{0_n})^2 \tag{1.13}$$

where N_b is the total number of covalent bonds, b_n is the distance between two atoms, b_{0_n} is the equilibrium distance and k_{b_n} is the force constant which is determined by comparing the experimental data after a conformational search

strategies. The two last parameters depend on the type of atoms i and j and the force fields.

1.3.1.2 Bond-angle bending

The bending potential captures the energy of the valence angle deformations between three atoms i, j and k joined by covalent bonds. This term is also calculated as an harmonic potential penalizing bond angles distant from the equilibrium value θ_0:

$$V^{angle} = \sum_{angles} \frac{k_\theta}{2}(\theta_{ijk} - \theta_0)^2 \tag{1.14}$$

where θ is the angle between three atoms i, j and k, θ_0 is the reference angle and k_θ is the force constant.

1.3.1.3 Torsion term

The third energy term concerns the ϕ dihedral angle of two plans defined by three covalent bonds and involving four atoms. A dihedral angle with a value of 0 is called *cis* and a dihedral angle with a value of 180 is called *trans*. The corresponding potential is defined as a Fourier series expansion around the equilibrium dihedral angle (Pitzer's potential):

$$V^{dihedral} = \sum_{torsions} \frac{V_n}{2}[1 + cos(n\phi_n - \gamma_n)] \tag{1.15}$$

where V_n is the rotational energy barrier, n is the number of minima in a complete rotation and γ is the phase angle.

A similar equation is used to calculate the energetic contribution of the improper dihedral angle potential.

1.3.2 The nonbonded interactions

The interaction between non-bonded atoms are described by the sum of two different energetic potentials (Equation 1.16). The non-bonded term is in principle calculated for all pairs of atoms but a number of pair-wise interactions are generally excluded using a cutoff.

$$V^{nonbonded} = \sum_{pairs} \left(V^{Lennard-Jones} + V^{Electrostatic} \right) \tag{1.16}$$

1.3.2.1 Lennard-Jones Potential

Van der Waals interactions are described by a Lennard-Jones Potential:

$$V^{Lennard-Jones} = 4\varepsilon_{ij} \sum_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{1.17}$$

where r_{ij} is the distance between the two atoms i and j, the ε (kcal.mol^{-1}) and σ (Å) are dependent on the atom type i and j.

The first term characterizes the repulsion between two atoms due to Pauli's exclusion principle while the second is an attractive term capturing the London dispersion forces. The Lennard-Jones potential is calculated for atoms separated by at least three covalent bonds. For atoms at exactly three covalent bonds from each other (interactions 1-4), the van der Waals energy is frequently divided by two.

1.3.2.2 Electrostatic Potential

As for the van der Waals interactions, the electrostatic potential is calculated between charged atoms separated by at least three covalent bonds. The potential is given by the Coulomb′s law:

$$V^{Electrostatic} = \sum_{ij} \frac{q_i q_j}{\pi 4 \varepsilon r_{ij}} \tag{1.18}$$

where q_i and q_j are the partial atomic charges of the atom i and j, ε is the effective dielectric constant and r_{ij} is the distance between the two atoms i and j.

1.4 Molecular Mechanics & Dynamics

MD is an intuitive method to explore the potential energy surface of a protein. Historically, one of the first molecules of interest, the small protein bovine pancreatic trypsin inhibitor, which has been studied by MD was in 1977.[67]

This technique is still broadly used in many different fields. The following section aims to give an overview of the methodology.

1.4.1 Energy minimization

Prior to perform a MD simulation, an energy minimization step of the potential energy of the system (1.6) is often essential to avoid too large molecular forces in the starting protein conformation that may encourage the exploration of low probability conformations. Usually, such high forces cause the numerical integration of the equations of motions to crash. Several algorithms can be used to perform a potential energy minimization such as the steepest descent (SD), conjugate gradient or quasi-newtonian method. Only the first approach has been used in this thesis. As with the conjugate gradient, SD is a gradient method using the potential energy $V(\mathbf{r})$ and its derivative with respect to $\mathbf{r}$.

$$\Delta r_i \approx \frac{\partial V(\mathbf{r}_i)}{\partial \mathbf{r}_i} \tag{1.19}$$

The system is almost exclusively 'pushed down' to reach the nearest local minimum on the potential energy surface. The conformation $\mathbf{r}(t_{n+1})$ at the n+1 step of the minimization step is computed by calculating the forces $\mathbf{f}(t_n)$ according to equation 1.7 for a given set of atomic coordinates $\mathbf{r}(t_n)$.

$$\mathbf{r}(t_{n+1}) = \mathbf{r}(t_n) + \frac{\mathbf{f}(t_n)}{\mid max(\mathbf{f}(t_n)) \mid} h_n \tag{1.20}$$

where h_n is the maximum displacement and f_n is the force at the step n. $\mid max(f_n) \mid$ is the maximum of the absolute values of the force components.

Then, the potential energy and the forces are calculated for the new atomic positions at the step $n + 1$.
The new conformation is accepted if $V_{n+1} < V_n$ and $h_{n+1} = 1.2h_n$. Otherwise, the new atomic positions are rejected and $h_n = 0.2h_n$. Calculations stop when the maximum of the absolute values of the force components (gradient) is smaller than a specified value or if the predefined maximum number of minimization steps has been achieved. While energy minimization approaches are effective at finding the nearest local energetic minimum from a starting conformation, they are unable to overcome energetic barriers. For this reason, MD simulations provide a more efficient way to sample low energy protein conformations.

1.4.2 Molecular dynamics

MD simulations are today widely used in order to study the structure, dynamics, and some thermodynamic aspects of molecular systems. Several algorithms, such as the leap-frog or Verlet integration, are implemented in GROMACS for integrating Newton's equations of motion (Equations 1.7, 1.9 and 1.10). Only the leap-frog method was used in the context of this thesis.

1.4.2.1 Integrators

In this section, a quick description of the Leap-frog and Verlet algorithms is given. The first one is based on the difference of two Taylor series of the velocity

$\mathbf{v}_i(t_n - \Delta t/2)$ and $\mathbf{v}_i(t_n + \Delta t/2)$ at a time $t = t_n$:

$$\mathbf{v}_i(t_n + \Delta t/2) = \mathbf{v}_i(t_n - \Delta t/2) + \frac{\Delta t}{m_i}\mathbf{f}_i(t_n) \tag{1.21}$$

Likewise, the atomic positions at a time t_n are determined as:

$$\mathbf{r}_i(t_n + \Delta t_n) = \mathbf{r}_i(t_n) + \Delta t \mathbf{v}_i(t_n + \frac{\Delta t_n}{2}) \tag{1.22}$$

The above two equations are used to integrate the equations of motion over the time and generate a trajectory. In order to get the position-update relation using the Verlet integrator, velocities $\mathbf{v}_i$ of the Equations 1.21 and 1.22 has to be removed. In addition, the time t_n has to be replaced by $t_n - \Delta t$ in Equation 1.21. Therefore, the new atomic positions are given by:

$$\mathbf{r}_i(t_n + \Delta t_n) = 2\mathbf{r}_i(t_n) - \mathbf{r}_i(t_n - \Delta t)\Delta t + \Delta t^2 \frac{\mathbf{f}_i(t_n)}{m_i} \tag{1.23}$$

The simulation time is directly dependent on the integration step Δt. The time step has to be smaller than the fastest motion of the system to be able to capture it. Rotations of hydrogen-bonded hydroxyl groups limit usually the time step in the order of the femtosecond.[68]

In a standard MD simulation, the total energy E of the molecular system is constant. Generally, the total number of atoms (N) and the volume (V) of the simulation box are also fixed. This type of simulation is called microcanonical or NVE simulation. However, in order to perform simulations under conditions that are the most similar to experiments, it may be better to maintain a constant temperature rather than energy (NVT simulations or canonical) and a constant pressure instead of the volume (NPT or isothermal-isobaric). For NVT and NPT

simulations, it is necessary to use a thermal or/and a pressure bath respectively.

1.4.2.2 Temperature Coupling

Several methods for controlling the temperature during MD simulations have been developed.[69] In the case of the Berendsen thermostat, the system is weakly coupled to a heat bath.[70]

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau_T} \tag{1.24}$$

The control of the temperature can be achieved by modifying the velocities of the particles of the system using a velocity rescaling factor λ (Equation 1.25).

$$\lambda^2 = 1 + \frac{\Delta t}{\tau}(\frac{T_0}{T} - 1) \tag{1.25}$$

where Δt is the integration step of the MD simulation and τ is the coupling constant. τ can be adjusted in function of the system. This constant has to be strong enough to maintain the average temperature of the system at reference value T_0 but without perturbing the dynamics.Usually for a time step of 3 fs, a coupling constant between 0.5 and 1.5 ps is sufficient. This thermostat suppresses fluctuations of the kinetic energy of the system. Therefore, the conformational sampling is biased and the produced trajectories are not consistent with the canonical ensemble. Recently, the Berendsen thermostat was modified using a stochastic procedure to yield canonical ensembles.[71] All simulations discussed in this thesis were performed using this stochastic Berendsen thermostat. In this velocity rescaling thermostat, an external stochastic term is used to correct the

kinetic energy distribution:

$$dK = (K_0 - K)\frac{dt}{\tau} + 2\sqrt{\frac{K_0 K}{N_f}}\frac{dW}{\sqrt{\tau}} \tag{1.26}$$

where K is the kinetic energy, N_f is the number of degrees of freedom and dW a Wiener noise.

1.4.2.3 Pressure Coupling

Different barostats can be used to maintain the simulation at constant pressure.[70, 72–74] They act by modifying the vectors of the simulation box and rescaling the atom coordinates. Three different methods are available:

- **isotropic pressure coupling:** modifications are applied uniformly to the system.
- **semi-isotropic pressure coupling:** modifications are applied independently in the x-y and z dimensions.
- **anisotropic pressure coupling:** modifications are applied independently in all directions.

In the major part of the thesis, simulations were performed in implicit solvent. Therefore, it was not necessary to define a simulation box and pressure coupling is not needed. As an example, the Berendsen barostat is described below. The term added to the equations of motion for maintaining a constant pressure is similar to the temperature coupling:

$$\frac{dP}{dt} = \frac{P_0 - P}{\tau_p} \tag{1.27}$$

where τ_p is the pressure coupling constant. The pressure can be expressed as a function of the kinetic energy of the system and the virial.

$$P = \frac{2}{3}\frac{K - w}{V} \tag{1.28}$$

where V is the volume of the simulation box and w is the virial defined as:

$$w = -\frac{1}{2}\sum_{\alpha<\beta}^{N_M} r_{\alpha\beta}\mathbf{F}_{\alpha\beta} \tag{1.29}$$

where $r_{\alpha\beta}$ is the distance between the center of mass of the molecules α and β at the time t and $\mathbf{F}_{\alpha\beta}$ is the force acting on the center of mass of the molecule α induced by the molecule β. Because the control of the pressure at constant temperature is linked to the volume by the isothermal compressibility κ_T, the pressure coupling is performed by adjusting the atomic coordinates and also the size of the simulation box with a correction factor μ (Equation 1.30).

$$\mu = [1 - \kappa_T\frac{\Delta t}{\tau_p}(P_0 - P)]^{\frac{1}{3}} \tag{1.30}$$

The Parrinello-Rahman barostat was also used in this thesis in the chapter 3.[72] This anisotropic approach allows to change the vectors of the simulation box but also its shape. Even if this method is slower than the Berendsen weak coupling, it maintains canonical ensemble and is more adapted to predict thermodynamic properties.

1.5 Enhanced Conformational Sampling

Most physical and chemical properties of a system can be interpreted directly or indirectly from free energy changes in the system. For example, the conformational preferences of a molecule or a protein, the solvation constants of association or dissociation of complexes are directly related to the difference in free energy between two states.

1.5.1 Free Energy Calculations

The free energy is a thermodynamic state function. For a system with a volume V that is defined by N particles at a temperature T, the Helmholtz free energy is given by:

$$F = E - TS \tag{1.31}$$

where E is the total energy of the system and S is the entropy.
In the same way, at constant pressure (P), the Gibbs's free energy is expressed as:

$$G = F + PV = E + PV - TS = H - TS \tag{1.32}$$

where H is the enthalpy.

According to statistical mechanics, the free energy for a discrete states system is directly related to the partition function Q:

$$Q = \sum_{i}^{allstates} \exp\left(-\frac{H_i}{\kappa_B T}\right) \tag{1.33}$$

where κ_B is the Boltzmann consant.

$$F = -\kappa_B T \ln Q \tag{1.34}$$

$$G = -\kappa_B T V \left(\frac{\partial \ln Q}{\partial V} \right)_T - \kappa_B T \ln Q \tag{1.35}$$

When those equations are applied in Cartesian coordinates to a system made up of a continuous number of microstates:

$$Q = \iint dp^N dr^N \exp\left(-\frac{H(p^N, r^N)}{\kappa_B T} \right) \tag{1.36}$$

$$F = -\kappa_B T \ln \left(\iint dp^N dr^N \exp\left[-\frac{H(p^N, r^N)}{\kappa_B T} \right] \right) \tag{1.37}$$

Q is a statistical property directly related to the probability of finding the system of interest in a given state and the ensemble of the phases accessible to the system. To obtain a good estimate of the absolute free energy, it would be necessary to sample the entire ensemble of conformations and calculate the partition function of the system. However, this approach is not possible because the system can actually adopt a very large number of conformations while all simulations or experiments lead to the sampling of a finite number of phases. The free energy difference between two states of the system is easier to calculate. It is obtained by the ratio of the probability of finding the system in the two considered states. Several simulation techniques have been designed to compute relative free energy changes such as thermodynamic integration, free energy perturbation or umbrella sampling.[75]

When one is interested in a reaction or physicochemical processes associated with a reaction coordinate s, it is interesting to compute the probability of the conformations observed for different values of s. s is also known as a collective variable (CV) and can describe any aspect of the system such as the distance between two center of mass, a dihedral angle or a helical structure. The probability distribution of the system along s is given by:

$$\rho(s) = \frac{\int \delta(s(\mathbf{r}) - s) \exp(-\frac{V(\mathbf{r})}{\kappa_B T}) d\mathbf{r}}{\int \exp(-\frac{V(\mathbf{r})}{\kappa_B T}) dr} \tag{1.38}$$

$$\rho(s) = \frac{\mathcal{Q}^p(s)}{\mathcal{Q}} \tag{1.39}$$

where $\mathcal{Q}$ is the partition function of the system and $V(\mathbf{r})$ is the potential energy of a given conformation of the system. The term $\mathcal{Q}^\rho(s)$ is actually the partition function of the system where all the degrees of freedom of s are fixed at a constant value. The probability density function can then be rewritten as a function of F:

$$\rho(s) = \exp\left(-\frac{F(s) - F}{\kappa_B T}\right) \tag{1.40}$$

where F is the free energy and $F(s)$ is the partial free energy according to s or Landau's free energy (Equation 1.41)

$$F(s) = -\kappa_B T \ln \mathcal{Q}^\rho(s) \tag{1.41}$$

Partial free energy is a function of s, which is directly related to the probability of sampling conformations for a specific value of s. In practice, Equation 1.40 is used to calculate the free energy (Equation 1.42).

$$\mathcal{F}(s) = -\kappa_B T \left(\ln p(s) - \ln \mathcal{Q}\right) \tag{1.42}$$

Because $\mathcal{Q}$ is independent of s, the probability density function $p(s)$ can be back-computed from a MD simulation by calculating the histogram of the collective variable along the simulation.

1.5.2 Umbrella Sampling

Quite frequently, classical MD simulations are unable to sample a significant range of values of a collective variable. The intramolecular potential or the system environment can induce constraints such that only a small part of the domain of variation of the reaction coordinate is explored. In addition, the conformational sampling is also limited since the simulation time is not infinite. Equation 1.42 does not allow to obtain directly the variation of free energy between the equilibrium state and a conformation of interest. The method of umbrella sampling (US) introduces an external energetic term to the potential energy of the system according to a CV.[76] In general, the bias is defined as a quadratic function:

$$V(\mathbf{r}; s)^{bias} = \frac{k}{2}(s - s_0)^2 \tag{1.43}$$

where k is the force constant and s_0 is the equilibrium position of the bias. Simulations performed using such methodologies are called biased simulations. By selecting different s_0 values, one is able to sample specific regions of the conformational space described by the CV. A simulation corresponding to a given value of s_0 is called a window and the free energy surface along the CV is called the potential of mean force (Figure 1.4).

In order to obtain the unbiased probability density of the system ($P(s)^u$), it is then

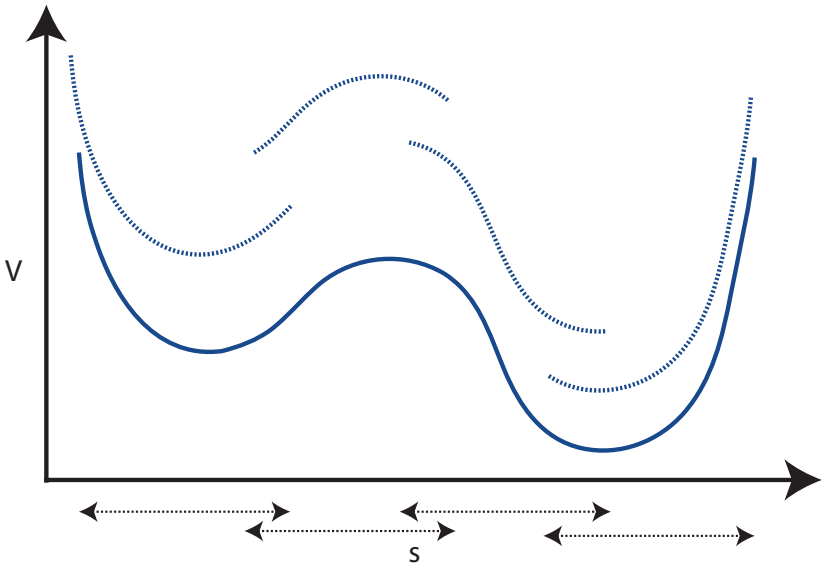

Figure 1.4: Example of umbrella sampling simulations. The real potential energy surface is depicted in blue (plain line). The potential energy surface corresponding to each window of s are represented with dashed lines.

necessary to remove the contribution of the bias to the computed probabilities. The unbiased distribution is given by equation 1.38.

However, US simulations provide only the biased distributions along the CV:

$$P(s)^b = \frac{\int \delta(s(\mathbf{r}) - s) \exp(-\frac{V(\mathbf{r})+V(\mathbf{r};s)^{bias}}{\kappa_B T}) d\mathbf{r}}{\int \exp(-\frac{V(\mathbf{r})+V(\mathbf{r};s)^{bias}}{\kappa_B T}) d\mathbf{r}} \tag{1.44}$$

Because the bias potential depends only on the collective variable s, the previous equation can be expressed as:

$$P(s)^b = \exp\left(-\frac{V(\mathbf{r};s)^{bias}}{\kappa_B T}\right) \frac{\int \delta(s(\mathbf{r}) - s) \exp(-\frac{V(\mathbf{r})}{\kappa_B T}) d\mathbf{r}}{\int \exp(-\frac{V(\mathbf{r})+V(\mathbf{r};s)^{bias}}{\kappa_B T}) d\mathbf{r}} \tag{1.45}$$

Using equation 1.38:

$$P(s)^u = P(s)^b \exp\left(\frac{V(\mathbf{r};s)^{bias}}{\kappa_B T}\right) \frac{\int \exp(-\frac{V(\mathbf{r})+V(\mathbf{r};s)^{bias}}{\kappa_B T}) d\mathbf{r}}{\int \exp(-\frac{V(\mathbf{r})}{\kappa_B T}) d\mathbf{r}} \tag{1.46}$$

$$P(s)^u = P(s)^b \exp\left(\frac{V(\mathbf{r};s)^{bias}}{\kappa_B T}\right) \left\langle \exp\left(-\frac{V(\mathbf{r};s)^{bias}}{\kappa_B T}\right) \right\rangle \tag{1.47}$$

At constant pressure, the free energy along s can be calculated as follows

$$G(s) = -\kappa_B T \ln P(s)^{bias} - V(\mathbf{r};s)^{bias} + F_i \tag{1.48}$$

Where F_i is a constant defined by

$$F_i = -\kappa_B T \ln \left\langle \exp\left(-\frac{V(\mathbf{r};s)^{bias}}{\kappa_B T}\right) \right\rangle \tag{1.49}$$

To be efficient, it is necessary to define accurately the range of each windows. Therefore, US is mainly relevant when the system of interest is well known and a substantial amount of experimental data are available. Otherwise, techniques such as metadynamics may be more attractive.

1.5.3 Metadynamics

The fundamental idea of metadynamics is to prevent a system from revisiting a part of the conformational space that has been already explored. The algorithm allows eliminating the problem of rare event sampling and to reconstruct the

multidimensional free energy profile of complex systems by introducing an history-dependent bias potential in a MD simulation defined as a small Gaussian:

$$V_G(s(x),t) = \omega \sum_{t'=\tau_G,2\tau_G,\ldots,t'<t} \exp\left(-\frac{[s(x)-s(x_G(t')]^2}{2\delta\sigma^2}\right) \tag{1.50}$$

where ω is the height and $\delta\sigma$ the width of the Gaussians, τ_G is the rate of their deposition, $s(x)$ and $s(x(t'))$ are the value of the collective variable.

One of the advantages of metadynamics is the possibility to fill up rapidly a local minimum and allow the exploration of the next lowest-energy minimum (Figure 1.5). The value of ω, $\delta\sigma$ and τ_G are crucial in this method because they influence directly the efficiency but also the accuracy of simulations.

The forces derived from the non-Markovian potential V_G act directly on the cartesian coordinates of the system, in addition to the forces f_{V_i} exerted by the potential energy V. During a metadynamics simulation, the force applied on an atom i of the system is given by:

$$f_i = f_{V_i} - \omega\frac{dt}{\tau_G}\sum_{t'\leq t}\left(\frac{s^t - s^{t'}}{2\delta\sigma^2}\right)\exp\left(-\frac{[s(x)-s(x_G(t')]^2}{2\delta\sigma^2}\right)\frac{\partial s(x)}{\partial x} \tag{1.51}$$

The historical potential V_G penalizes visited areas of the conformational space encouraging the system to explore new states. V_G allows to speed up the simulation of rare events. The system escapes from a local minimum through the nearest transition state encountered. The ability to reconstruct the free energy profile from metadynamics simulations is based on the assumption that $F_G(s,t) = -V_G(s,t)$ is an approximation of $F(s)$ in the region sampled by $s(x_G(t'))$. In this way, after

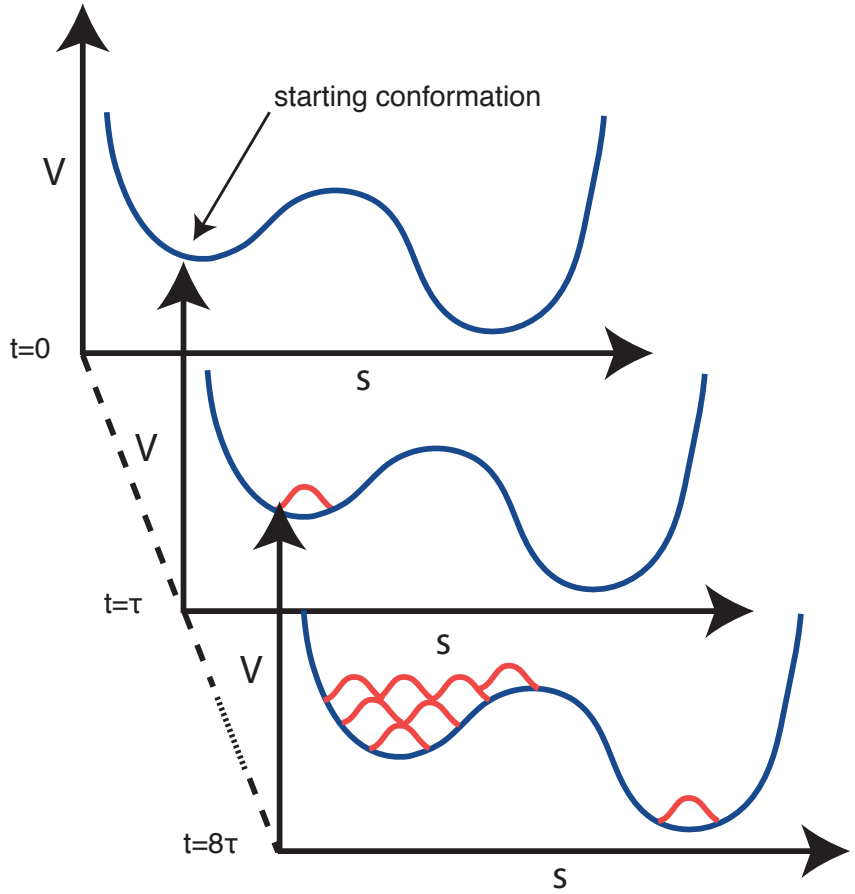

Figure 1.5: Illustration of a metadynamics simulation.The potential energy surface of the protein is represented in blue. The gaussian added every τ time steps are depicted in red. This figure illustrates how the system can easily escape from a local minimum and explore other conformations of interest.

a sufficient time when the CV has converged, the free energy profile is given by:

$$\lim_{x \to +\infty} F_G(s,t) \sim F(s) \tag{1.52}$$

Metadynamics is particularly efficient to explore the conformational space of system by biasing the simulation using up to three CVs. For more complex systems, other variants such as bias exchange metadynamics may be preferred.[77]

1.5.4 Bias-Exchange metadynamics

In this thesis, the bias-exchange variant of metadynamics was also used.[77] The approach entails running a set of molecular dynamics simulations. The sampling of molecular conformations in each simulation is biased by a history-dependent potential as described above. Exchanges between the biasing potentials used in the different CVs are periodically attempted according to a replica exchange scheme. The swap is accepted according to the Metropolis criterion. BEMD has been shown to allow exploring complex free energy landscape such as the folding free energy landscape of small proteins and protein/ligand complexes on timescales of a few dozen ns.[78, 79]

1.5.5 Weighted Histogram Analysis Method

When a single window was defined to bias a system along a CV, equation 1.48 is sufficient to reconstruct the free energy profile (FEP). Otherwise, the FEP is reconstructed within a constant (F_i) and computational approaches may be used

to connect profiles obtained from different windows (Figure 1.4). The constants F_i (equation 1.49) correspond to the free energies associated with the introduction of the biased potential $V(\mathbf{r};s)^{bias}$. In this thesis, the Weighted Histogram Analysis Method (WHAM) was used to reconstruct the FEP from biased simulations along a CV (s) depending on the cartesian coordinates r of the system.[80, 81] The overall distribution $P^u(s)$ is obtained by calculating a weighted average of the distributions of each window, minimizing the statistical error.

Considering N_w as the number of biased simulations and n_i as the number of stochastically independent events used to construct the biased distribution $P_i^b(s)$, the overall unbiased distribution $P^u(s)$ can be expressed according to the unbiased distributions $P_i^u(s)$ of each window as:

$$P^u(s) = \sum_{i=1}^{N_w} P_i^u(s) \times \frac{n_i \exp[-\beta(V(\mathbf{r};s)_i^{bias}(s) - F_i)]}{\sum_{j=1}^{N_w} n_j \exp[-\beta(V(\mathbf{r};s)_j^{bias}(s) - F_j)]} \tag{1.53}$$

Using Equation 1.47, this expression can be simplified as

$$P^u(s) = \sum_{i=1}^{N_w} \frac{n_i P_i^b(s)}{\sum_{j=1}^{N_w} n_j \exp[-\beta(V(\mathbf{r};s)_j^{bias}(s) - F_j)]} \tag{1.54}$$

The constants F_j can be directly estimated from the unbiased distribution $P^u(s)$:

$$\exp(-\beta F_j) = \int P^u(s) ds \exp(-\beta V(\mathbf{r};s)_j^{bias}(s)) ds \tag{1.55}$$

At each step of the process, the Equations 1.54 and 1.55 have to be solved. The WHAM approach can be easily generalized to more than one CV.

1.5 Bibliography

[1] Daniel A Dias, Sylvia Urban, and Ute Roessner. A historical overview of natural products in drug discovery. *Metabolites*, 2(2):303–336, March 2012.

[2] David J Newman and Gordon M Cragg. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *Journal of Natural Products*, 75(3):311–335, March 2012.

[3] Jörg Eder, Richard Sedrani, and Christian Wiesmann. The discovery of first-in-class drugs: origins and evolution. *Nature Reviews Drug Discovery*, 13(8):577–587, August 2014.

[4] David Brown and Giulio Superti-Furga. Rediscovering the sweet spot in drug discovery. *Drug Discovery Today*, 8(23):1067–1077, December 2003.

[5] Andrew M Davis, Simon J Teague, and Gerard J Kleywegt. Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angewandte Chemie International Edition*, 42(24):2718–2736, June 2003.

[6] Jesse W-H Li and John C Vederas. Drug discovery and natural products: end of an era or an endless frontier? *Science*, 325(5937):161–165, July 2009.

[7] Joseph G Lombardino and John A Lowe. A guide to drug discovery: The role of the medicinal chemist in drug discovery — then and now. *Nature Reviews Drug Discovery*, 3(10):853–862, October 2004.

[8] Joseph A DiMasi, Ronald W Hansen, and Henry G Grabowski. The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, 22(2):151–185, March 2003.

[9] Michael Dickson and Jean Paul Gagnon. The Cost of New Drug Discovery and Development. *Discovery Medicine*, 4(22):172–179, June 2009.

[10] John P Overington, Bissan Al-Lazikani, and Andrew L Hopkins. How many drug targets are there? *Nature Reviews Drug Discovery*, 5(12):993–996, December 2006.

[11] Chu Qin, Cheng Zhang, Feng Zhu, Feng Xu, Shang Ying Chen, Peng Zhang, Ying Hong Li, Sheng Yong Yang, Yu Quan Wei, Lin Tao, and Yu Zong Chen. Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Research*, 42(D1):D1118–D1123, January 2014.

[12] Peter Imming, Christian Sinning, and Achim Meyer. Drugs, their targets and the nature and number of drug targets. *Nature Reviews Drug Discovery*, 5(10):821–834, October 2006.

[13] Mathias Rask-Andersen, Surendar Masuram, and Helgi B Schiöth. The druggable genome: Evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annual Review of Pharmacology and Toxicology*, 54:9–26, January 2014.

[14] Konrad H Bleicher, Hans-Joachim Böhm, Klaus Müller, and Alexander I Alanine. Hit and lead generation: beyond high-throughput screening. *Nature Reviews Drug Discovery*, 2(5):369–378, May 2003.

[15] Valère Lounnas, Tina Ritschel, Jan Kelder, Ross McGuire, Robert P Bywater, and Nicolas Foloppe. Current progress in Structure-Based Rational Drug Design marks a new mindset in drug discovery. *Computational and Structural Biotechnology Journal*, 5:e201302011, 2013.

[16] Mark A Lindsay. Innovation: Target discovery. *Nature Reviews Drug Discovery*, 2(10):831–838, October 2003.

[17] Eugene C Butcher, Ellen L Berg, and Eric J Kunkel. Systems biology in drug discovery. *Nature Biotechnology*, 22(10):1253–1259, October 2004.

[18] Mark A Lindsay. Finding new drug targets in the 21st century. *Drug Discovery Today*, 10(23-24):1683–1687, December 2005.

[19] Ricardo Macarron, Martyn N Banks, Dejan Bojanic, David J Burns, Dragan A Cirovic, Tina Garyantes, Darren V S Green, Robert P Hertzberg, William P Janzen, Jeff W Paslay, Ulrich Schopfer, and G Sitta Sittampalam. Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery*, 10(3):188–195, March 2011.

[20] Nicolas Moitessier, Pablo Englebienne, Devin Lee, Janice Lawandi, and Christopher R Corbeil. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *British Journal of Pharmacology*, 153 Suppl 1:S7–26, March 2008.

[21] Antonio Lavecchia and Carmen Di Giovanni. Virtual screening strategies in drug discovery: a critical review. *Current Medicinal Chemistry*, 20(23):2839–2860, November 2013.

[22] William L Jorgensen. The many roles of computation in drug discovery. *Science*, 303(5665):1813–1818, March 2004.

[23] James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British Journal of Pharmacology*, 162(6):1239–1249, February 2011.

[24] Andrew L Hopkins, György M Keserü, Paul D Leeson, David C Rees, and Charles H Reynolds. The role of ligand efficiency metrics in drug discovery. *Nature Reviews Drug Discovery*, 13(2):105–121, February 2014.

[25] Diane Joseph-McCarthy, Arthur J Campbell, Gunther Kern, and Demetri Moustakas. Fragment-based lead discovery and design. *Journal of Chemical Information and Modeling*, 54(3):693–704, March 2014.

[26] Michael J Keiser, Vincent Setola, John J Irwin, Christian Laggner, Atheir I Abbas, Sandra J Hufeisen, Niels H Jensen, Michael B Kuijer, Roberto C Matos, Thuy B Tran, Ryan Whaley, Richard A Glennon, Jérôme Hert, Kelan L H Thomas, Douglas D Edwards, Brian K Shoichet, and Bryan L Roth. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, November 2009.

[27] Christopher A Lipinski. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, 1(4):337–341, December 2004.

[28] Paul D Leeson and Brian Springthorpe. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews Drug Discovery*, 6(11):881–890, November 2007.

[29] Tudor I Oprea and Hans Matter. Integrating virtual screening in lead discovery. *Current Opinion in Chemical Biology*, 8(4):349–358, August 2004.

[30] Craig Umscheid, David Margolis, and Craig Grossman. Key Concepts of Clinical Trials: A Narrative Review. *Postgraduate medicine*, 123(5):194–204, September 2011.

[31] Amy C Anderson. The process of structure-based drug design. *Chemistry & Biology*, 10(9):787–797, September 2003.

[32] Ana L Carvalho, José Trincão, and Maria J Romão. X-ray crystallography in drug discovery. *Methods in Molecular Biology*, 572:31–56, October 2009.

[33] Lasse Jenner, Sergey Melnikov, Nicolas Garreau de Loubresse, Adam Ben-Shem, Madina Iskakova, Alexandre Urzhumtsev, Arturas Meskauskas, Jonathan Dinman, Gulnara Yusupova, and Marat Yusupov. Crystal structure of the 80S yeast ribosome. *Current Opinion in Structural Biology*, 22(6):759–767, December 2012.

[34] Maurizio Pellecchia, Ivano Bertini, David Cowburn, Claudio Dalvit, Ernest Giralt, Wolfgang Jahnke, Thomas L James, Steve W Homans, Horst Kessler, Claudio Luchinat, Bernd Meyer, Hartmut Oschkinat, Jeff Peng, Harald Schwalbe, and Gregg Siegal. Perspectives on NMR in drug discovery: a technique comes of age. *Nature Reviews Drug Discovery*, 7(9):738–745, September 2008.

[35] Pierre Montaville and Nadège Jamin. Determination of membrane protein structures using solution and solid-state NMR. *Methods in Molecular Biology*, 654(4):261–282, June 2010.

[36] Jens Carlsson, Ryan G Coleman, Vincent Setola, John J Irwin, Hao Fan, Avner Schlessinger, Andrej Sali, Bryan L Roth, and Brian K Shoichet. Ligand discovery from a dopamine D3 receptor homology model and crystal structure. *Nature chemical biology*, 7(11):769–778, November 2011.

[37] Alexander Wlodawer and Jiri Vondrasek. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annual Review of Biophysics and Biomolecular Structure*, 27:249–284, June 1998.

[38] Rebecca C Wade. 'Flu' and structure-based drug design. *Structure*, 5(9):1139–1145, September 1997.

[39] Louis A Clark, P Ann Boriack-Sjodin, John Eldredge, Christopher Fitch, Bethany Friedman, Karl J M Hanf, Matthew Jarpe, Stefano F Liparoto, You Li, Alexey Lugovskoy, Stephan Miller, Mia Rushe, Woody Sherman, Kenneth Simon, and Herman Van Vlijmen. Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. *Protein Science*, 15(5):949–960, May 2006.

[40] Kamil Khafizov, Carlos Madrid-Aliste, Steven C Almo, and Andras Fiser. Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *Proceedings of the National Academy of Sciences of the United States of America*, 111(10):3733–3738, March 2014.

[41] Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, 3(8):711–715, August 2004.

[42] Ricardo Macarron. Critical review of the role of HTS in drug discovery. *Drug Discovery Today*, 11(7-8):277–279, April 2006.

[43] Thomas H Keller, Arkadius Pichota, and Zheng Yin. A practical view of 'druggability'. *Current Opinion in Chemical Biology*, 10(4):357–361, August 2006.

[44] Andrew L Hopkins and Colin R Groom. The druggable genome. *Nature Reviews Drug Discovery*, 1(1):727–730, September 2002.

[45] Andreas P Russ and Stefan Lampel. The druggable genome: an update. *Drug Discovery Today*, 10(23-24):1607–1610, December 2005.

[46] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 46(1-3):3–26, March 2001.

[47] Daniel F Veber, Stephen R Johnson, Hung-Yuan Cheng, Brian R Smith, Keith W Ward, and Kenneth D Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, 45(12):2615–2623, June 2002.

[48] Oleg Ursu, Anwar Rayan, Amiram Goldblum, and Tudor I Oprea. Understanding drug-likeness. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5):760–781, April 2011.

[49] Giulio Vistoli, Alessandro Pedretti, and Bernard Testa. Assessing drug-likeness – what are we missing? *Drug Discovery Today*, 13(7-8):285–294, April 2008.

[50] Alan C Cheng, Ryan G Coleman, Kathleen T Smyth, Qing Cao, Patricia Soulard, Daniel R Caffrey, Anna C Salzberg, and Enoch S Huang. Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnology*, 25(1):71–75, January 2007.

[51] Robert P Sheridan, Vladimir N Maiorov, M Katharine Holloway, Wendy D Cornell, and Ying-Duo Gao. Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *The Journal of Chemical Physics*, 50(11):2029–2040, November 2010.

[52] Fredrik N B Edfeldt, Rutger H A Folmer, and Alexander L Breeze. Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discovery Today*, 16(7-8):284–287, April 2011.

[53] Philip J Hajduk, Jeffrey R Huth, and Stephen W Fesik. Druggability indices for protein targets derived from NMR-based screening data. *Journal of Medicinal Chemistry*, 48(7):2518–2525, April 2005.

[54] Tom Halgren. New method for fast and accurate binding-site identification and analysis. *Chemical Biology & Drug Design*, 69(2):146–148, February 2007.

[55] Thomas A Halgren. Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling*, 49(2):377–389, February 2009.

[56] Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, 10:168–168, January 2009.

[57] Peter Schmidtke and Xavier Barril. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *Journal of Medicinal Chemistry*, 53(15):5858–5867, August 2010.

[58] Jesus Seco, F Javier Luque, and Xavier Barril. Binding site detection and druggability index from first principles. *Journal of Medicinal Chemistry*, 52(8):2363–2371, April 2009.

[59] Ahmet Bakan, Neysa Nevins, Ami S Lakdawala, and Ivet Bahar. Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence

of Probe Molecules. *Journal of Chemical Theory and Computation*, 8(7):2435–2447, July 2012.

[60] Olgun Guvench and Alexander D MacKerell. Computational fragment-based binding site identification by ligand competitive saturation. *PLoS Computational Biology*, 5(7):e1000435, July 2009.

[61] Simon J Teague. Implications of protein flexibility for drug discovery. *Nature Reviews Drug Discovery*, 2(7):527–541, July 2003.

[62] Yechun Xu, Jacques-Philippe Colletier, Martin Weik, Hualiang Jiang, John Moult, Israel Silman, and Joel L Sussman. Flexibility of aromatic residues in the active-site gorge of acetylcholinesterase: X-ray versus molecular dynamics. *Biophysical Journal*, 95(5):2500–2511, September 2008.

[63] Israel Silman and Joel L Sussman. Acetylcholinesterase: 'classical' and 'non-classical' functions and pharmacology. *Current Opinion in Pharmacology*, 5(3):293–302, June 2005.

[64] Terrone L Rosenberry. *Acetylcholinesterase*. Advances in Enzymology and Related Areas of Molecular Biology. John Wiley & Sons, Inc., Hoboken, NJ, USA, November 2006.

[65] Todd F Kagawa, Jakki C Cooney, Heather M Baker, Sean McSweeney, Mengyao Liu, Siddeswar Gubba, James M Musser, and Edward N Baker. Crystal structure of the zymogen form of the group A Streptococcus virulence factor SpeB: an integrin-binding cysteine protease. *Proceedings of the National Academy of Sciences of the United States of America*, 97(5):2235–2240, October 1999.

[66] Daniel Alvarez-Garcia and Xavier Barril. Relationship between Protein Flexibility and Binding: Lessons for Structure-Based Drug Design. *Journal of Chemical Theory and Computation*, 10(6):2608–2614, June 2014.

[67] J Andrew McCammon, Bruce R Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267:585–590, June 1977.

[68] Alexey K Mazur. Hierarchy of fast motions in protein dynamics. *The Journal of Physical Chemistry B*, 102(2):473–479, January 1998.

[69] Philippe H Hünenberger. Thermostat algorithms for molecular dynamics simulations. *Advanced Computer Simulation*, 173:105–149, January 2005.

[70] Herman J C Berendsen, Johan P M Postma, Wilfred F van Gunsteren, Alfredo Di Nola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *Journal of Chemical Information and Modeling*, 81(8):3684–3690, June 1984.

[71] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, January 2007.

[72] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182, December 1981.

[73] Shuichi Nosé and Michael L Klein. Constant pressure molecular dynamics for molecular systems. *Molecular Physics*, 50(5):1055–1076, August 1983.

[74] Glenn J Martyna, Mark E Tuckerman, Douglas J Tobias, and Michael L Klein. Explicit reversible integrators for extended systems dynamics. *Molecular Physics*, 87(5):1117–1157, May 1996.

[75] Christophe Chipot, M Scott Shell, and Andrew Pohorille. Free Energy Calculations. *Theory and Applications in Chemistry and Biology*, 86(18):33–72 119–167, June 2007.

[76] Johannes Kästner. Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(6):932–942, May 2011.

[77] Stefano Piana and Alessandro Laio. A Bias-Exchange Approach to Protein Folding. *The Journal of Physical Chemistry B*, 111(17):4553–4559, May 2007.

[78] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5):826–843, September 2011.

[79] Vanessa Leone, Fabrizio Marinelli, Paolo Carloni, and Michele Parrinello. Targeting biomolecular flexibility with metadynamics. *Current Opinion in Structural Biology*, 20(2):148–154, April 2010.

[80] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8):1011–1021, October 1992.

[81] Marc Souaille and Benoit Roux. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Computer Physics Communications*, 135(1):40–57, March 2001.

2

Protein-Ligand Interactions

This chapter gives an overview of the protein-ligand binding process with a focus on the specific case of intrinsically disordered proteins.

According to quantum mechanics, a system of one or more particles is defined by its wave function Ψ (also called state function). The Schrödinger equation determines how the wave function evolves over time:[1]

$$i\hbar\frac{\partial\Psi}{\partial t} = \hat{H}\Psi \tag{2.1}$$

where i is the imaginary unit, $\hbar$ is the reduced Planck constant and $\hat{H}$ is the Hamiltonian operator.

In the case of a stationary system, which is not explicitly time-dependent, the Schrödinger equation can be expressed as:

$$E\Psi = \hat{H}\Psi \tag{2.2}$$

where E is the energy of the stationary state.

In the case of a molecular system composed of M nuclei and n electrons, the Hamiltonian of the system is calculated using the following equation:

$$\begin{aligned}\hat{H} = &-\sum_{i=1}^{M}\frac{\hbar^2}{2m_i}\nabla_i^2 - \sum_{j=1}^{n}\frac{\hbar^2}{2m_j}\nabla_j^2 + \frac{e^2}{2}\sum_{i=1}^{M}\sum_{k\neq i}^{M}\frac{z_i z_k}{|\mathbf{r}_i - \mathbf{r}_k|} \\ &+ \frac{e^2}{2}\sum_{j=1}^{n}\sum_{j\neq l}^{n}\frac{1}{|\mathbf{r}_j - \mathbf{r}_l|} - e^2\sum_{i=1}^{M}\sum_{j=1}^{n}\frac{z_i}{|\mathbf{r}_i - \mathbf{r}_j|}\end{aligned} \tag{2.3}$$

where $\mathbf{r}$ are the coordinates of nuclei and electrons, z is the charge the nucleus, e is the charge of an electron, m is the mass and ∇ is the Laplace operator.

Several approaches have been introduced in order to provide an accurate description of small systems within reasonable time.[1] However, they are computationally demanding and may require very important resources even for systems of a few atoms. Therefore, as described in the previous chapter, interactions between atoms and molecules are frequently modeled using a classical formalism and different equations are needed to model different aspects of quantum chemistry. This chapter gives an overview of the protein-ligand binding process with a focus on the specific case of intrinsically disordered proteins (IDPs).

2.1 Non-covalent interactions

The functioning of biological systems is based on folding and recognition mechanisms involving non-covalent molecular interactions. At the protein level, a subtle balance between attractions and repulsions controls the three-dimensional (3D) structure of a protein and therefore also its activity in the cell. In this part, three kind of non-covalent interactions are discussed.

2.1.1 Electrostatic Interactions

The electrostatic interactions between two charged molecules can be described by Coulomb′s law:

$$V^{Electrostatic} = \frac{q_i q_j}{\pi 4 \varepsilon r_{ij}} \tag{2.4}$$

where q_i and q_j are the partial atomic charges of the atom i and j, ε is the effective dielectric constant and r_{ij} is the distance between the two atoms i and j.

This electrostatic potential is inversely proportional to the distance between the two charges. It is important to note that the relative permittivity of water is about 80 at room temperature, which means that the ionic interactions are considerably reduced in aqueous medium compared to air ($\varepsilon = 1$) causing the dissolution of most salt crystals in water.

The equation 2.4 can be generalized to the case of protein-ligand interactions:

$$V_{P-L}^{Electrostatic} = \sum_{i=1}^{N_P} \sum_{j=1}^{N_L} \frac{q_i q_j}{\pi 4 \varepsilon r_{ij}} \tag{2.5}$$

where N_P and N_L are respectively the number of partial atomic charges of the protein and the ligand.

2.1.1.1 Van der Waals Interactions

Interactions between neutral molecules are based on electrostatic interactions between permanent dipoles and/or induced dipoles. These forces are responsible for multiple interactions between neighboring atoms and are also called van der Waals forces. We can distinguish three kind of van der Waals interactions: Keesom force, Debye force and London dispersion force.

Keesom force When, in a neutral molecule, the center of gravity of the positive charges is different from the center of gravity of the negative charges, the molecule is considered as polar and has an electric dipole moment μ directed

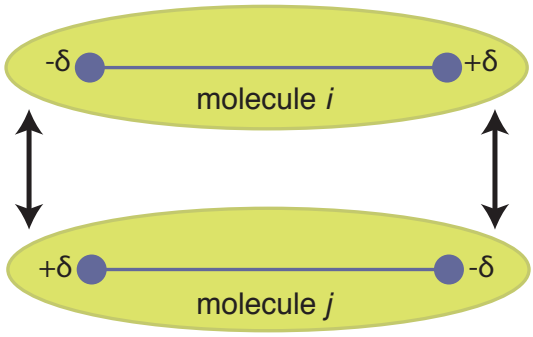

Figure 2.1: Illustration of the interaction between two permanent dipoles.

from the negative charge to the positive charge. Two polar molecules of non-zero dipole moments can find favorable positions to maximize the attraction between them (Figure 2.1).

The energy corresponding to dipole-dipole interactions is calculated as follows:

$$V^{Keesom} = \frac{-2\mu_i^2\mu_j^2}{3\left(4\pi\varepsilon_0\varepsilon_r\right)^2\kappa_B T r_{ij}^6} \tag{2.6}$$

where μ is the dipole moment, ε_0 is the permittivity of free space, ε_r is the dielectric constant of surrounding material, κ_B is the Boltzmann constant and T is the temperature.

Debye force A polar molecule with a permanent dipole moment (μ) induces a rearrangement of the electron cloud of neighboring apolar molecules under the effect of the electric field ($\mathbf{E}$). The electron cloud deformation is characterized by the polarizability of the molecule, which increases with the number of electrons. This non polar molecule acquires an induced dipole moment:

$$\mu_i = \alpha\mathbf{E} \tag{2.7}$$

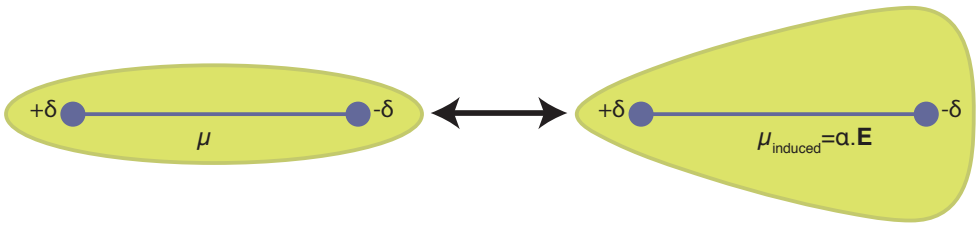

Figure 2.2: Illustration of the interaction between a permanent dipole (μ) and an induced dipole (μ_{induced}).

where α is the polarizability of the apolar molecule.

This induced dipole interacts with the permanent dipole of the first molecule as shown in figure 2.2.

The energy corresponding to dipole-dipole induced interactions is expressed as:

$$V^{Debye} = \frac{4\mu_i^2\alpha}{(4\pi\varepsilon_0\varepsilon_r)^2 r_{ij}^6} \tag{2.8}$$

London dispersion force In the case of non polar molecules, the electron cloud is symmetrically distributed and no dipole moment is observed. However, the electron motion may create an instantaneous dipole moment able to polarized neighboring apolar molecules leading to an induced dipole moment. Both instantaneous dipoles vary rapidly over time and can interact together as shown in Figure 2.3.

The London dispersion is described the following equation:

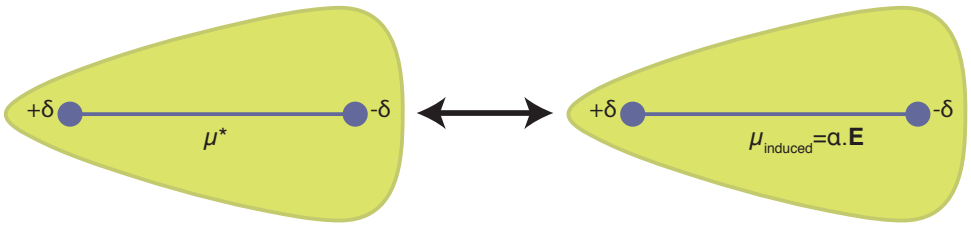

Figure 2.3: Illustration of the interaction between an instantaneous dipole (μ^*) and an induced dipole (μ_{induced}).

$$V^{London} = \frac{2\alpha_i\alpha_j}{3\left(4\pi\varepsilon_0\varepsilon_r\right)^2 r_{ij}^6}\frac{I_iI_j}{I_i + I_j} \tag{2.9}$$

where I is the ionization potentials of the molecule.

The London forces are very weak interactions. However, the large number of interatomic contacts in protein-ligand complexes make that dispersion forces may play an important role in the binding process.

Van der Waals radius and energy By considering only van der Waals forces as attractive, it is not possible to explain the existence of an equilibrium intermolecular distance. Repulsive forces are also involved in the formation of protein-ligand complexes controlling the impenetrability of the molecules. Van der Waals interactions are described by a Lennard Jones potential:

$$V^{Lennard-Jones} = \varepsilon_{ij}(\frac{\sigma_{ij}}{r_{ij}})^{12} - 2\varepsilon_{ij}(\frac{\sigma_{ij}}{r_{ij}})^6 \tag{2.10}$$

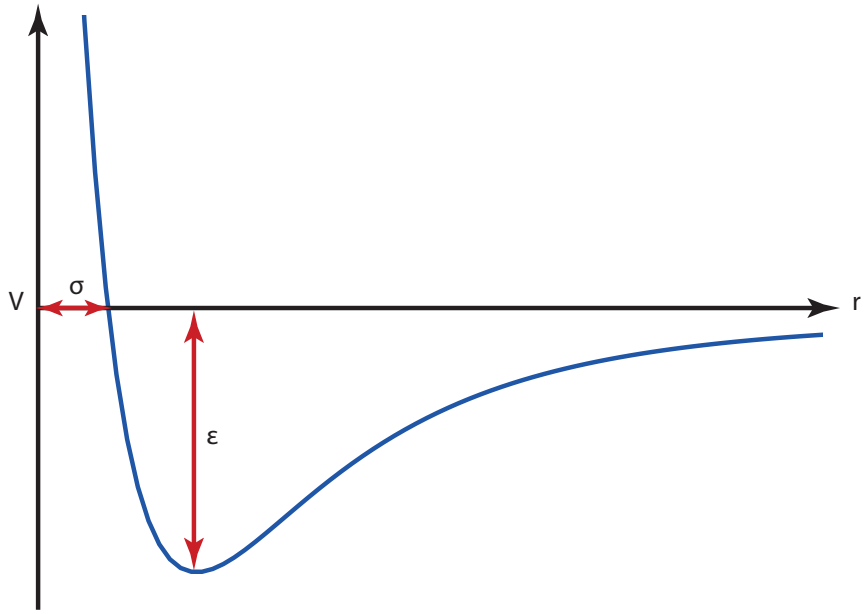

Figure 2.4: Illustration of a Lennard-Jones potential (Equation 2.10).

where r_{ij} is the distance between the two atoms i and j, the ε (kJ.mol^{-1}) and σ (Å) are dependent on the atom type i and j.

This potential is illustrated in figure 2.4.

Van der Waals interactions are weak interactions (around several kJ.mol^{-1}) but can significantly stabilize a protein-ligand complex.

2.1.1.2 π interactions

π interactions are mainly due to dispersion forces.[2, 3] Different kind of π interactions can be observed such as:

1. **cation-π system**: interactions between the positive charge of cations (or a metal) and the face of a π system

2. **Polar-π system**: interactions of a polar molecule and the multipole moment of a π system

3. π **stacking**: interactions between two aromatic systems ('face-to-face')

Non-covalent interactions involving π systems are very important in many biological events such as protein-ligand recognition.[2]

2.1.2 Hydrogen-bond

Hydrogen-bonds are also mainly electrostatic interactions. They involve dipole/induced-dipole interactions even if other phenomena such as polarization or dispersion also contribute to the total energy of hydrogen bonding. During this process, a hydrogen atom attached to an electronegative atom (donor) is carrying a fraction of positive charge that polarize another molecule with a lone pair (acceptor). These electronegative atoms are usually fluorine, oxygen, or nitrogen. A hydrogen attached to a carbon atom may also participate in hydrogen bonding if the carbon atom is bound to electronegative atoms. The strong interaction between the dipole and the induced dipole involves the alignment of atoms as is the case for van der Waals forces. Hydrogen bonding is highly directional and usually stronger than van der Waals interactions (between 10 to 30 kJ.mol^{-1}). Nevertheless all hydrogen bonds do not have the same characteristics. As suggested by Jeffrey, they may be classified into three different categories according to their binding energy and directionality (Table 2.1).[4]

	Strong Interactions	Medium Interactions	Weak Interactions
Kind of Interactions	Mostly Covalent	Mostly Electrostatic	Electrostatic London
Bond Energy kJ.mol^{-1}	60-160	15-60	< 15
Bond Lengths (r) Å	1.2-1.5	1.5-2.2	2.2-3.2
Bond Angles (θ)	$175 - 180°$	$130 - 180°$	$90 - 150°$

Table 2.1: Main characteristics of the different kind of hydrogen-bonds.[4]

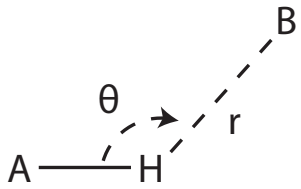

Figure 2.5: Representation of a hydrogen bond according to the parameter r and θ given in Table 2.1 .

2.1.3 Hydrophobic Interactions

Hydrophobic interactions play a crucial role in many biological processes such as protein folding or assembly of biological membranes.[5–7] Water is characterized by a strong internal cohesion that is characterized by a high enthalpy of vaporization and a high surface tension. Therefore, an apolar molecule tends to avoid contacts with water by interacting with the non polar parts of other molecules. This association partially offsets the unfavorable free energy due to the solvation of such molecules by reducing the area accessible to water and creating strong van der Waals interactions.

The hydrophobic effect is a complex process that is still not well characterized. Hydrophobic interactions appear to be related to the occurrence of

transient dipoles, water molecule rearrangements.[8–10] Despite their apolarity, the electron clouds of two adjacent hydrophobic molecules interact in such a way that partial charges of opposite sign appear. Therefore, London forces are the main process characterizing the hydrophobic effect even if other parameters are also involved.[11]

2.2 Thermodynamics of ligand binding

In the context of this thesis, a molecular receptor is a protein providing a structural arrangement of its functional groups promoting interactions with another molecule called a ligand. These interactions may lead to the formation of a reversible complex because no covalent bonds are formed. In some cases, the receptor conformations are so specific that favourable interactions can only be formed with a few ligands (principle of selectivity). The formation of a reversible receptor-ligand complex is usually represented as a chemical equilibrium. This state is reached when the concentrations of both reactants and products do not change over time. Actually, the equilibrium reflects a compensation between the reaction rates of the forward and backward reactions. A simplified view of the formation of a complex involving only one protein receptor (R) and one ligand (L) can be expressed as follows:

$$R + L \rightleftharpoons RL \tag{2.11}$$

This chemical process is often characterized by the dissociation constant of the complex or K_d depending on the concentrations of each species, at equilibrium:

$$K_d = \frac{[R][L]}{[RL]} \tag{2.12}$$

The lower the K_d is, the higher is the affinity between the ligand and its receptor. K_d is directly related to the standard Gibbs free energy of binding ΔG^0 by:

$$\Delta G^0 = RT \ln \frac{K_d}{C^0} \tag{2.13}$$

where R is the gas constant, T is the temperature and C^0 is the standard state concentration of a dilute solute (1 mol.L^{-1}).
A negative value of the ΔG^0 means that the free energy of the complex is lower than the free energies of each partner in an unbound state. Therefore, the formation of the complex is spontaneous. This free energy can also be expressed as the sum of two terms:

$$\Delta G^0 = \Delta H - T\Delta S^0 \tag{2.14}$$

where ΔH is the variation of enthalpy and ΔS^0 is the variation of entropy.

The enthalpy change in this process is related to changes in non-covalent interactions. If $\Delta H < 0$, the system is considered more stable because the bound state involves more interactions (or fewer but stronger) than the free state. The entropy change captures if the system becomes more ordered ($\Delta S^0 < 0$) or less ordered ($\Delta S^0 > 0$) after the formation of the complex. In the context of protein-ligand binding, this property is usually associated with the solvation entropy change, the protein/ligand conformational entropy changes, and the protein/ligand rotational and translational entropy changes.[12] If an increase of the disorder is observed, the reaction is entropically favorable. Indeed, if $\Delta S^0 > 0$

and $\Delta H < 0$, then ΔG^0 could be negative. In general, the presence of the ligand stabilizes the protein and ΔS^0 becomes negative. If this entropic cost is not offset by a decrease of the enthalpy term, ΔG^0 will be close to 0. This characteristic allows proteins to be involved in rapid association/dissociation processes.

The affinity characterizing receptor-ligand interactions is a subtle balance between the entropy term and the enthalpy term. Electrostatic interactions play a fundamental role in the stability of the complex. They include salt bridges, hydrogen bonds, π-π interactions, dipole-dipole interactions and also interactions with metallic ions. Hydrogen bonds are due to the attraction of a hydrogen atom bonded to an electronegative atom (donor) by another electronegative atom or a π-electron system (acceptor). The electronegative atoms are usually fluorine (F), nitrogen (N) or oxygen (O). The distance between donor and acceptor atoms is between 2.5 Å and 3.2 Å, and the bond angle is between 130° and 180°. The strength of the hydrogen bond depends directly on the environment, and more especially on the dielectric constant ε. The dielectric constant of water (or the relative permittivity), as well as that found at the protein surface, is estimated at 80. However, inside the protein, ε is evaluated between 1 and 20.[13] Furthermore, the dielectric constant near polar groups and flexible regions is higher than in apolar regions. Thus, in the context of ligand binding, hydrogen bonds buried in the protein are generally more important than those exposed to the solvent. Before the binding process, protein and ligand are only interacting with the solvent. In solution, the functional groups of each species are involved in hydrogen bonding with water molecules. The difference between the free energies of these contributions and the hydrogen bonds formed in the complex determines whether these hydrogen bonds contribute favorably to the formation of the complex or not. Indeed, the presence of polar groups in the protein/ligand molecules that are not involved in hydrogen bonding in the complex is highly unfavourable to complex

formation.[14] In contrast, the desolvation of apolar parts releases highly ordered water molecules increasing the entropy. This increase in entropy of the solvent due to the burial of the protein apolar regions is called the hydrophobic effect.[5] The hydrophobic effect is generally the major force that stabilizes the complex, while the Coulomb interactions and hydrogen bonds rather intervene in the specificity of receptor-ligand interactions. The buried hydrophobic surface can be correlated to the free energy of binding with values between -0.11 to -0.24 kJ.mol^{-1}.Å^{-2}. For example, the burial of a methyl group representing a surface of 25 Å^2 may correspond to a contribution of -2.75 to -6 kJ.mol^{-1}. Hydrophobic interactions are responsible for roughly 80% of the free energy involved in molecular recognition events.[15] Beside the increase in the solvent entropy, the binding process involves a decrease of the solute entropy. The change in free energy due to the loss of side-chain conformational entropy ($T\Delta S$) was found to vary from 0 (alanine, glycine, proline) to 8.7 kJ.mol^{-1} (glutamine) with an average value of 3.7 kJ.mol^{-1}.[16]

The knowledge of the enthalpy and the entropy terms allow a better understanding of the interactions compared to the dissociation constant. Indeed, processes having similar ΔG^0 may have very different ΔH and ΔS^0. Carbonic anhydrase II is a good example illustrating this problem. The enzyme catalyses the transformation of carbon dioxide in water to bicarbonate with the release of protons. Several small molecules are known to inhibit carbonic anhydrase II such as 4-carboxybenzene-sulfonamide and 5-dimethylamino-1-naphthalene-sulfonamide. The chemical structures and the binding thermodynamic properties of both ligands are represented in Figure 2.6.[17]Interestingly, those two ligands show a similar free energy of binding and thus a comparable dissociation constant. However, different enthalpic and entropic components are observed. In both cases, the enthalpy is favourable whereas from an entropy point of view, the first interaction is unfavourable while the second is favourable. This analysis reflects

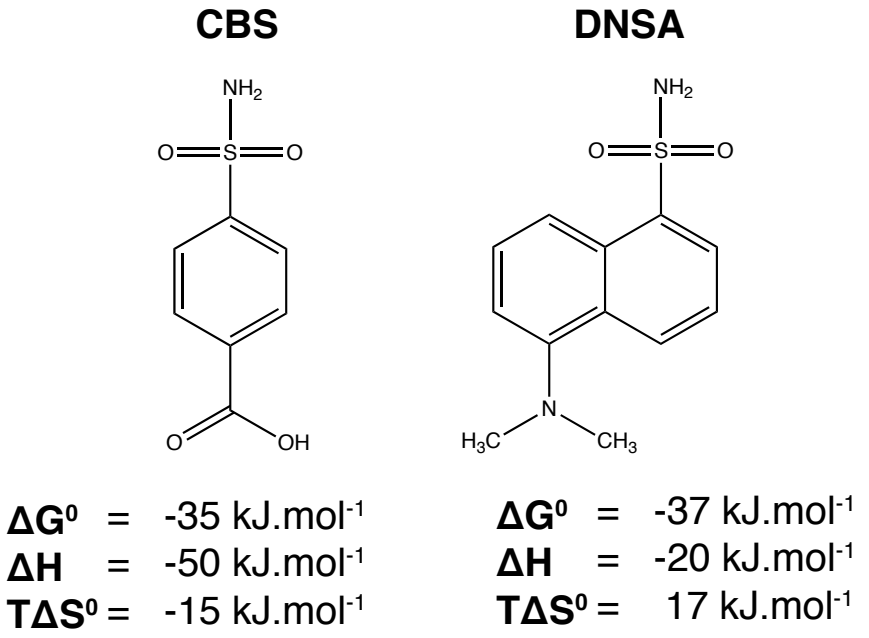

Figure 2.6: Chemical structure of two different inhibitors of carbonic anhydrase II with different thermodynamic properties determined by isothermal titration calorimetry at 298 K. CBS: 4-carboxybenzene-sulfonamide, DNSA: 5-dimethylamino-1-naphthalene-sulfonamide.[17]

that the driving forces leading to the complex formation are different.

2.3 Measuring binding free energies

As described in the previous section, the logarithm of the dissociation constant is proportional to the Gibbs free energy of binding (Equation 2.13). Several methodologies, such as isothermal titration calorimetry or surface plasmon resonance, are commonly used to experimentally measure K_d and other thermodynamic properties.[18, 19] Such approaches allow to measure K_d values between 10^{-2}

to 10^{-10} M, which corresponds to values of the free energy between -10 and -70 kJ.mol^{-1} at a temperature of 298 K. A change in the free energy of 5.7 kJ.mol^{-1} at 298 K induces a perturbation of the dissociation constant by a factor of ten. The dissociation constant is not the only parameter that can be measured to describe the affinity between a ligand and its target. K_i and IC_{50} characterize the inhibition of a protein. K_i represents, in the case of an enzyme for example, and more especially in the ideal conditions of the Michaelis-Menten model ($\Delta G^0 << 0$ and a concentration of substrate much larger than the concentration of product), the inhibition constant.[20] The half maximal inhibitory concentration , noted IC_{50} allows to characterize the effect of a small molecule on the biological activity of a target. When measuring those values, it is essential to perform experiments under such conditions that the target is a limiting factor, which means that the dynamic phenomena associated with the target must be a linear function of the concentration of the target. The inhibitor binds the enzyme either alone or to the enzyme interacting with its substrate, depending on whether the inhibitor is competitive or non-competitive. The dissociation constant at equilibrium is called K_i and corresponds to the concentration of inhibitor required to saturate half of the enzyme's active site. Therefore, K_i allows to measure the affinity of an inhibitor for a specific enzyme. The action of a competitive inhibitor on enzyme activity is measured by the IC_{50}. The binding process of a ligand to its receptor is similar to the fixation of a substrate to the enzyme, except that the binding phenomenon is not followed by a change in enzymatic activity. In the case of a reversible and non cooperative binding process, the Cheng-Prusoff relationships express the link between IC_{50} and K_i:[21]

$$K_i = \frac{IC_{50}}{1 + \frac{[S]}{K_m}} \tag{2.15}$$

where K_i is the binding affinity of the inhibitor, IC_{50} is the functional strength of the inhibitor, $[S]$ is the substrate concentration and K_m is the concentration of substrate at which enzyme activity is at half maximal (Michaelis-Menten constant).

2.4 Protein flexibility in ligand binding

In order to recognize specifically a ligand, a protein possesses a suitable binding site. Based on this observation, a first model explaining the mechanism of binding of a ligand to a protein was proposed by Fischer in 1894 (Figure 2.7 A).[22] In this approach so-called 'lock and key', proteins and ligands are considered as rigid bodies. Although the model is applicable to some extent to large number of complexes, it does not reflect the general behaviour of the different protagonists in solution. For this reason, the induced-fit model was introduced (Figure 2.7 B, left).[23] First, the ligand interacts with a conformation of the target. Then, during the binding process, each partner may adjust its structure to maximize the binding affinity. This mechanism dominated the view of of protein flexibility until a new vision of protein folding emerged during the 1990s.[24, 25] According to the rate-determining step of the binding process, several models such as the conformational selection or the population-shift mechanism were proposed.[26–28] Those approaches are based on the unbound state, which exists as a set of conformations called conformational isomers or conformers. The ligand can select one or few conformations from the equilibrium ensemble to form a complex (Figure 2.7 B, right).

A large number of systems where protein flexibility plays a crucial role in the

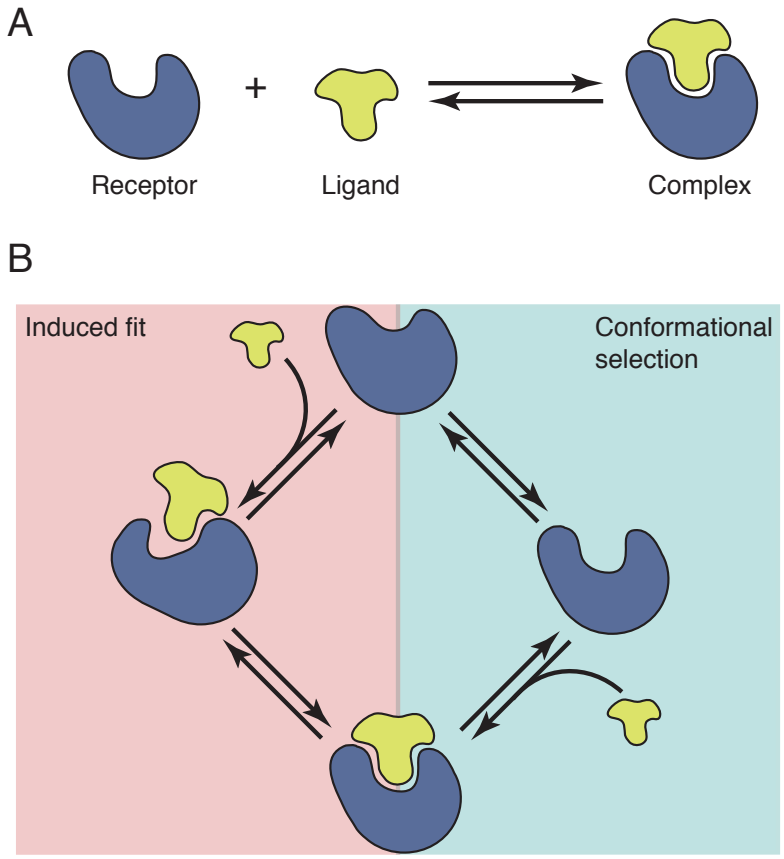

Figure 2.7: Illustration of the most common models describing the binding process between a protein (receptor) and a ligand. A) Lock and key model. B) Induced fit model (left panel) and Conformational selection (right panel).

binding process have been described in the literature.[29] This flexibility may involve small or large structural rearrangements. For instance, motions of few amino acids located near the active site of acetylcholinesterase have been reported.[30] Because of its involvement in the memorization process and Alzheimer's disease, this enzyme has been extensively studied.[31, 32] Using MD simulations, the authors were able to identify two residues of the active site showing a high flexibility. The findings from this study suggest that the equilibrium ensemble of this protein is composed of a large number of distinct conformations, and different ligands select a different subset of those conformations upon complex formation.[30]

Another interesting example is given by the antibody SPE7.[33] Free SPE7 exists in two very different conformations (Ab1 and Ab2). Those isomers are able to interact with two structurally different ligands forming two distinct complexes (Ab3 and Ab4). Initially, the binding mode appeared to follow the mechanism of the conformational selection. However, the authors suggest that it is possible to induce Ab4 starting from Ab1 and Ab3 starting from Ab2. This observation tips the scale in favor of the induced-fit model. This study highlights the potential role of conformational diversity in cross-reactivity leading to autoimmune diseases and allergies.[34]

2.5 Intrinsically Disordered Proteins

The structure and function relationships of proteins occupy a central and fundamental position in biology and have been studied extensively. For a long time, it has been accepted that the three-dimensional (3D) structure of a protein was only dictated by its amino acid sequence and also that this specific structure

was related to a single function.[36] However, in the 1990s, the discovery of proteins that are not or poorly ordered have led many to question this dogma. The development of spectroscopic techniques has accelerated the study of 3D structures of such disordered proteins. The idea that proteins could be active while being unstructured became increasingly stronger. Thus, in 1999, Dyson et al. suggested that a protein may be both partially (or completely) unstructured and active.[37] Since then, intrinsically disordered proteins (IDPs) have been studied extensively and a database called Disprot was created.[38] In 2014, 694 proteins and 1539 disordered regions were referenced. Proteins that contain a segment of at least 30 consecutive disordered residues in their native state are typically classified as IDPs (intrinsically disordered proteins).[39] In mammals, around 50% of the proteins can be considered as partially or completely disordered.[40] The considerable flexibility of IDPs facilitates interactions with a broad range of proteins and explains why IDPs often play key roles in important cellular processes such as signaling or transcription.[41, 42] In addition, IDPs are involved in many cancers, cardiovascular and neurodegenerative diseases.[43]

The flexibility of IDPs is due to their singular amino acid composition. Globular proteins are well structured, and usually composed of hydrophobic residues forming the core of the protein whereas polar and charged residues are more frequently localized at the protein surface. It is now well accepted that IDPs are significantly enriched in proline , glutamic acid, lysine, serine and glutamine while they are depleted in tryptophan, tyrosine, phenylalanine, cysteine, isoleucine, leucine, and asparagine.[44] Furthermore, they have a high net charge and low hydrophobicity, precluding the formation of a hydrophobic core and promoting instead an extended conformation by electrostatic repulsion between charged groups. In solution, by contrast with globular proteins, IDPs do not adopt one dominant structure, but oscillate between structurally diverse conformations of

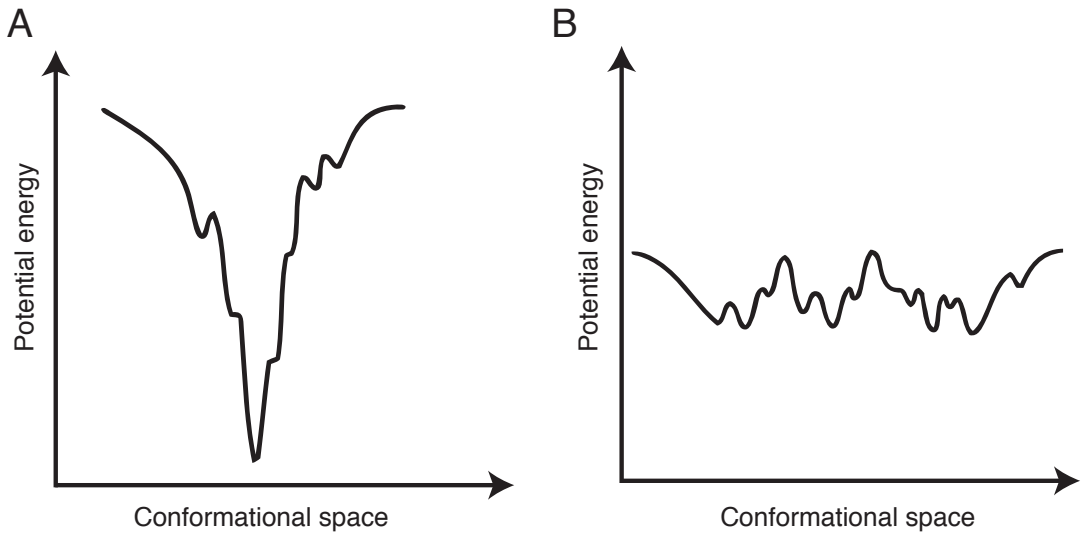

Figure 2.8: Illustration of the energy landscape of a globular protein (A) and an intrinsically disordered protein (B).

comparable low energies. This translates into a flat energy landscape (Figure 2.8).[45]

2.6 IDPs-ligand Interactions

IDPs have just started to be considered as druggable.[46] The chapter section is focused on the mechanisms of small-molecules binding to IDPs. Molecular recognition between an IDP and a partner protein can involve a disorder-to-order transition through a coupled folding upon binding mechanism, which produces high-specificity low-affinity complexes (Figure 2.9).[47–49] There are, however, several examples of IDPs that remain disordered upon complex formation.[50] IDPs are attractive therapeutic targets as they are often implicated in a broad range of diseases, such as cancers, cardiovascular disease or neurodegenerative

diseases. However, the considerable flexibility of IDPs presents a challenge for drug discovery approaches.[43] Owing to their lack of a well-defined tertiary structure, it is generally not possible to determine the structure of isolated IDPs. So far, structure-based approaches to inhibit IDPs have targeted either partner proteins that are ordered or ordered complexes, in those cases where IDPs fold upon binding. For instance, the p53 tumor suppressor is an IDP that is involved in the progression of more than 50% of human cancers.The transcriptional activity of p53 is tightly regulated by its partner protein MDM2 (murine double minute 2) and cancer cells often overexpress MDM2 to inhibit p53 function.[51] As the p53-binding domain of MDM2 is folded, crystal structures can be readily obtained and have been exploited to design several classes of small-molecule inhibitors of p53-MDM2.[52] Some of the most successful inhibitors have advanced into clinical trials.[53] However, several protein-protein interactions involve two IDPs whose structure cannot be solved in isolation. Even in those instances where two IDPs mutually fold upon binding, the structure of the complex may not reveal pockets to which small molecules could readily bind. Thus a more general route to inhibiting IDP function would be to directly target their disordered state with small molecules. Historically, this approach has not been considered feasible.[46] However, this view has been challenged in recent years, with the realization that several small molecules inhibit IDP function by binding to their unfolded state.[54–56] The interactions of small molecules with IDPs challenge our understanding of molecular recognition and it is important to clarify the mechanisms of IDP-small molecule interaction before such proteins can be more routinely targeted. This is here illustrated with a review of three well-studied systems: the oncoprotein c-Myc, Aβ (amyloid β-peptide) and α-synuclein.

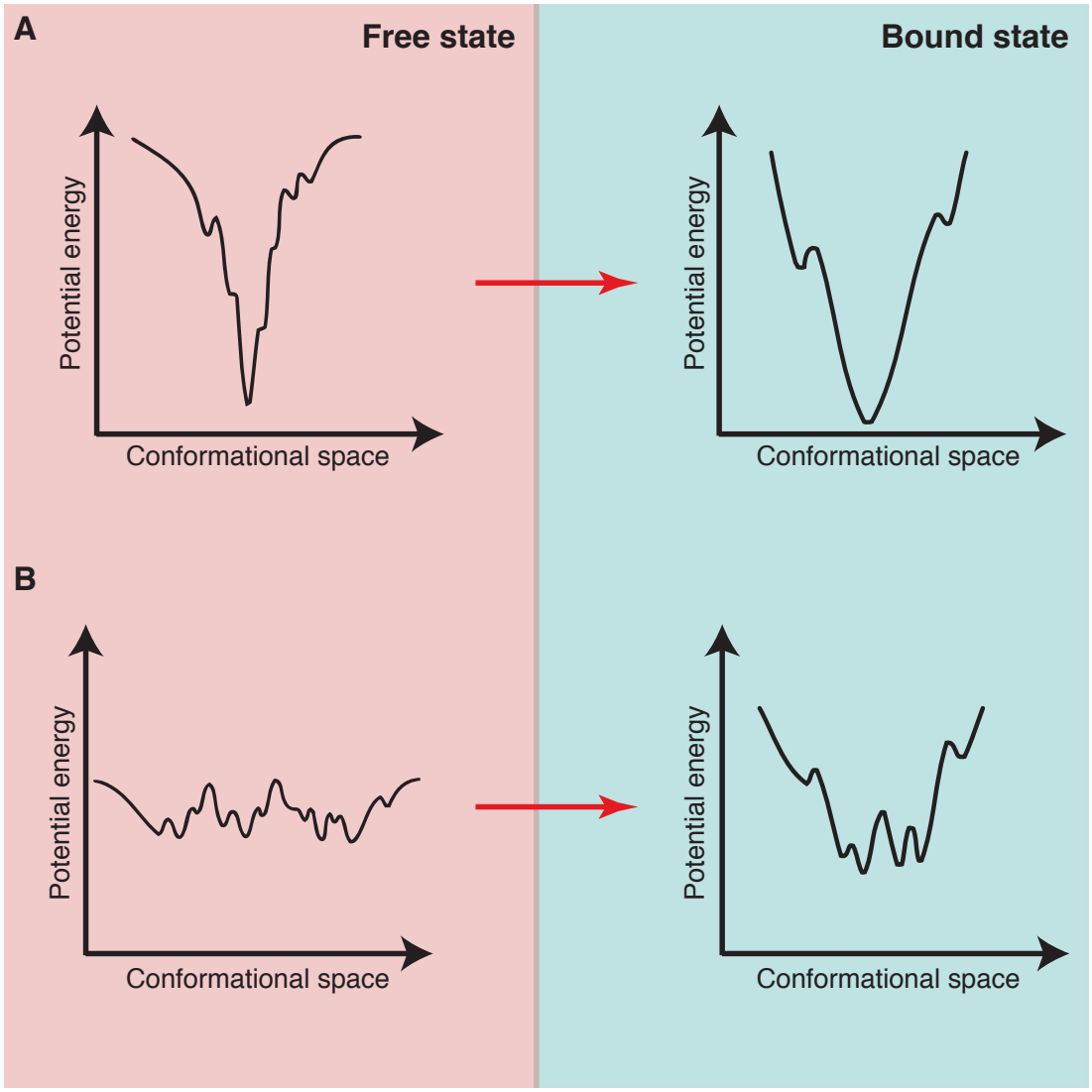

Figure 2.9: The impact of ligand binding on the energy landscape of a globular protein (A) and an intrinsically disordered protein (B).

2.6.1 c-Myc

The proto-oncogene protein c-Myc consists of 439 amino acids and contains an 88-amino-acid bHLHZip (basic helix-loop-helix leucine zipper) domain. In its monomeric form, c-Myc is intrinsically disordered.[57] c-Myc has been shown to interact with a large number of other proteins. The specific interaction between c-Myc and the protein Max has been studied extensively because the c-Myc-Max heterodimer binds DNA and regulates gene expression.[58] It has been shown that overexpression of c-Myc is frequent in many cancers, and disruption of the c-Myc-Max interaction is a possible anticancer strategy.[41] Structurally diverse small molecules inhibiting the formation of this complex were discovered through a yeast two-hybrid screen.[54] Biophysical studies using fluorescence assays, nuclear magnetic resonance (NMR) and circular dichroism (CD) measurements were performed to characterize protein-ligand interactions.[57, 59, 60] These studies suggest that the small molecules disrupt the c-Myc-Max interaction by stabilizing conformations in monomeric c-Myc that are incompatible with heterodimerization with Max. Three distinct binding sites, encompassing residues 366-375, 375-385 and 402-409, have been mapped on to the c-Myc bHLHZip domain.[59] Remarkably, the three distinct c-Myc-binding sites can be occupied simultaneously by different ligands. These results suggest that the c-Myc-small molecule interactions are fairly localized and can be predicted from primary sequence analysis. Indeed, protein disorder prediction algorithms can locate approximately the small molecule-binding sites of c-Myc, which tend to be enriched in hydrophobic amino acids in comparison with the rest of the domain.[57] In addition, many of the small-molecule ligands can bind truncated c-Myc segments containing a single binding site with a binding affinity similar to that of the full c-Myc bHLHPZip domain.

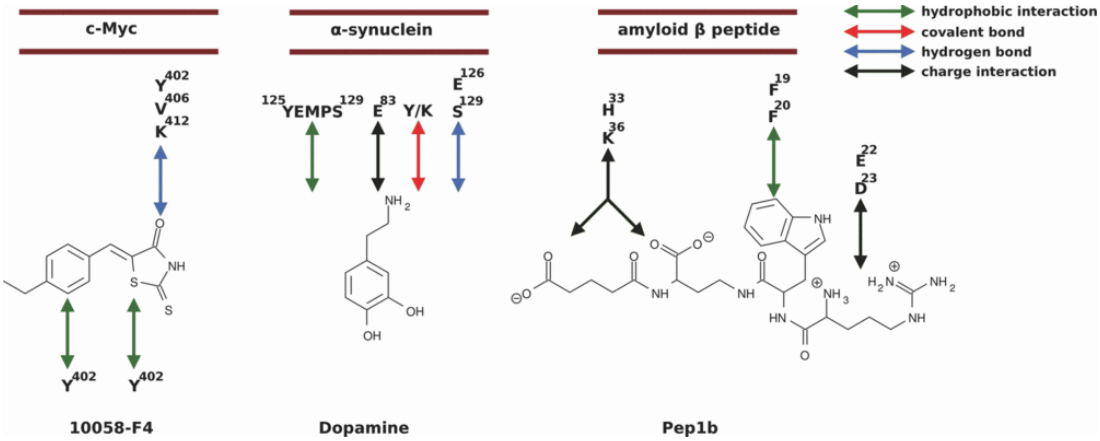

Figure 2.10: Summary of the main interactions observed in three IDP-ligand complexes: c-Myc-10058-F4, α-synuclein-dopamine and Aβ-Pep1b. The amino acids are represented using the one-letter-code.

For instance, the small molecule 10058-F4 binds in a fluorescence polarization assay to c-Myc353-437 with a K_d of 5.3±0.7 μM and to c-Myc402-412 with a K_d of 13.3±1 μM.[57] Furthermore, similar chemical shift perturbations were observed for c-Myc353-437 and c-Myc402-412 upon binding 10058-F4. NMR and CD studies suggest that c-Myc remains disordered upon binding 10058-F4. Ligand binding appears to lead to formation of a hydrophobic cluster between the ligand and the side chains of Tyr402, Ile403, Leu404 and Val406 (Figure 2.10). Molecular dynamics studies detailed in chapter 3 reveal multiple distinct binding modes for 10058-F4, with frequent stacking interactions with Tyr402 as well as hydrogen-bonding interactions with the backbone of Tyr402, Val406 and Lys412.[61]

2.6.2 Amyloid β-peptide

Alzheimer′s disease is a neurodegenerative pathology characterized by the formation of senile plaques in the brain.[62] The aggregation of Aβ is known to be one of the main components of those plaques and may be associated with the pathogenesis of Alzheimer's disease.[63, 64] Aβ (36-43 amino acids) is produced by the successive cleavage of the APP (amyloid precursor protein) by the enzymes β-secretase and γ-secretase. Although the role of APP is not completely characterized, it appears to be crucial for synapse formation and function.[65] The aggregation of Aβ, as well as with other compounds such as apoliprotein E, induces the development of senile plaques. Aβ adopts a folded helical structure in membrane environments, but an aggregation-prone β-sheet conformation in aqueous solution.[66] Over the last few decades, many peptide and small molecule inhibitors of Aβ aggregation have been discovered, primarily through in vitro assays.[67] Current small molecule inhibitors appear to inhibit Aβ aggregation through at least two distinct mechanisms. For instance, scylloinositol derivatives have been shown by electron microscopy experiments to bind and stabilize monomeric and trimeric forms, thus blocking aggregation.[68, 69] On the other hand, compounds such as Thioflavin T or Congo Red appear to interact with Aβ aggregates, although decades of studies on these compounds have produced several conflicting models of binding mechanisms. Plausible hypotheses have been recently reviewed extensively by Groenning.[70] Computational studies have attempted to clarify protein-ligand interactions. Molecular dynamics simulations were performed recently for ten small-molecule inhibitors in the presence of a truncated form of Aβ (Aβ12-28).[71] Although the small molecules did not exhibit a predominant binding mode and did not dramatically affect the secondary-structure

preferences of Aβ12-28, a number of conserved interactions with Aβ12-28 could be observed. Most of the ligands interacted preferentially with the N-terminal portion of the peptide (residues 13-20). Energetic analysis revealed favourable electrostatic interactions with three amino acids (His13, His14 and Lys16). Additionally, favourable hydrophobic interactions are observed between the inhibitors and the entire N-terminal stretch, with the sites of highest interaction probability being near the side chains of Phe19 and Phe20. The binding affinities appear to be roughly correlated with the number of aromatic groups and charged groups present in the ligands. Molecular dynamics simulations have also been performed to examine the interactions of two small ligands, Pep1b and Dec-DETA, that were designed to stabilize the central helix in Aβ.[72] Both ligands appear to stabilize the Aβ central helix (residues 15-24) in Aβ13-26 by interacting preferentially with two charged amino acids: Glu22 andAsp23. In addition, electrostatic interactions with His13 and Lys16 as well as hydrophobic interactions with Phe19 and Phe20 were also reported for Pep1b (Figure 2.10). It appears that the extended side-chain interactions between the ligands and Aβ disfavour intramolecular side-chain interactions that would destabilize the central α-helix. Recently, molecular dynamics simulations were used to study the interactions of inositol ligands with (Gly-Ala)4 modelled either as disordered or β-sheet aggregates of four peptides, or as an extended fibril-like oligomer.[73] The ligands were observed to form predominantly one or two hydrogen bonds with the peptide backbone. The results suggested that inositol does not inhibit amyloid formation by dispersing preformed aggregates or by preventing aggregation, but is more likely to bind instead to the surface of prefibrillar aggregates.[73] The computed dissociation constants of the ligands were two orders of magnitude higher than those measured experimentally, suggesting that additional sidechain interactions must contribute significantly to the binding affinity of the inositol ligands to Aβ aggregates.[73]

2.6.3 α-synuclein

The 140-amino-acid protein α-synuclein consists of three distinct domains. The central region of α-synuclein is known to be crucial for the aggregation of α-synuclein fibrils, one of the main components of Lewy bodies associated with many neurodegenerative diseases such as Parkinson's disease.[74, 75] Under physiological conditions, α-synuclein normally adopts a helical conformation that is non-pathogenic and plays a role in neurotransmitter release. It is still not well understood how α-synuclein first forms soluble oligomers called protofibrils, followed by the development of β-sheet-rich α-synuclein fibrils. In light of these observations, a deeper molecular-level understanding of interactions between monomeric, protofibril and fibril forms is important to facilitate the discovery of small molecule inhibitors of α-synuclein fibrillization. A few years ago, 15 fibrillization inhibitors were found by screening a small-molecule library using a fibrillization assay.[76] Many of these inhibitors are members of the catecholamine family and include dopamine. There is controversy about the mechanisms of interactions between dopamine and α-synuclein. Conway et al. have suggested that dopamine readily oxidizes into dopamine-derived orthoquinones that subsequently form a covalent adduct with α-synuclein by radical coupling to form dityrosine linkages or by nucleophilic attack of a lysine side chain.[76] On the other hand, Norris et al. failed to detect significant levels of dopamine-α-synuclein adducts and suggested instead that binding occurs through non-covalent interactions with the α-synuclein segment Tyr125-Glu-Met-Pro-Ser129.[77] Herrera et al. used docking calculations and molecular dynamics simulations to study the interactions of dopamine and several plausible oxidized derivatives with an NMR-derived structural ensemble of α-synuclein.[78] In the majority of the simulated complexes,

the ligands interacted through a broad range of hydrogen-bonding and hydrophobic interactions with the region Tyr125-Glu-Met-Pro-Ser129. Additionally, significant electrostatic interactions were computed between the ligands and Glu83 located in the non-β-amyloid region of α-synuclein. These predictions were tested by a series of biophysical experiments. Point mutations to alanine in the Tyr125-Glu-Met-Pro-Ser129 region did not prevent dopamine inhibition of α-synuclein aggregation in an in vitro fibrillization assay, suggesting that dopamine interacts nonspecifically with this region. On the other hand, mutation of Glu83 to alanine strongly impaired the ability of dopamine to inhibit α-synuclein aggregation.[78] Non-catecholamine inhibitors of α-synuclein aggregation have also been identified. A broad range of biophysical methods were used by Lendel et al. to characterize the interactions of Congo Red and lacmoid with α-synuclein.[79] They concluded that these two small molecules interact broadly with the N-terminal and central region of α-synuclein as small oligomeric species.[79]

2.6.4 Conclusion

Although small molecules have now been found to interact directly with several IDPs in their monomeric form, an important challenge is to clarify the specificity of the interactions. For instance, there are numerous proteins that contain a bHLHZip domain similar to that of c-Myc. Consequently, several small molecules that inhibit the c-Myc-Max complex also inhibit related bHLHZip pairs. To illustrate, the compound 10058-F4 has also been shown in a yeast two-hybrid assay to disrupt the complexes MyoD- E2-2, Mad1-Max and Mxi1-Max, although several other bHLHZip pairs were not inhibited.[54] Several of the dopamine derivatives that inhibit α-synuclein aggregation have also been shown

to also dissolve fibrils of Aβ in vitro.[80] Congo Red and lacmoid bind readily to α-synuclein, a protein closely related to α-synuclein which does not aggregate under physiological conditions.[79] In several cases, relatively structurally diverse small molecules have been found to interact with similar regions in an IDP. Additionally, many studies suggest that the complexes between small molecules and IDPs remain disordered.[81] This suggests that the binding of the small molecules is driven by a large number of weak interactions.[46] Arguably, unlike proteins, small molecules are unlikely to induce IDP folding upon binding, as the relatively limited intermolecular contacts that they form are unlikely to overcome the large conformational entropy loss necessary to structure an IDP. Structure-based approaches to design ligands for IDPs will therefore have to explicitly consider multiple binding modes. Although the mechanisms of IDP aggregation are still not well understood, a number of small-molecule inhibitors of IDP aggregation have reached clinical studies. For instance, methylthionium chloride, initially developed as an antimalarial agent, has been shown to inhibit in vitro the aggregation of the IDP tau.[82] Results of a Phase II clinical trial reported that methylthionium chloride slows down cognitive impairment in patients suffering from Alzheimer's disease, thus inhibiting the formation of tau aggregates is a promising strategy for the development of Alzheimer's disease treatments.[83]

2.6 Bibliography

[1] Alan Hinchliffe. Modelling molecular structures. *John Wiley & Sons*, July 2000.

[2] Emmanuel A Meyer, Ronald K Castellano, and François Diederich. Interactions with aromatic rings in chemical and biological recognition. *Angewandte Chemie (International ed. in English)*, 42(11):1210–1250, March 2003.

[3] Jiří Klimeš, David R Bowler, and Angelos Michaelides. Chemical accuracy for the van der Waals density functional. *Journal of physics. Condensed matter : an Institute of Physics journal*, 22(2):022201, January 2010.

[4] George A Jeffrey. An Introduction to Hydrogen Bonding. *Oxford University Press*, 1997.

[5] Walter Kauzmann. Some factors in the interpretation of protein denaturation. *Advances in protein chemistry*, 14:1–63, November 1959.

[6] Marvin Charton and Barbara I Charton. The structural dependence of amino acid hydrophobicity parameters. *Journal of theoretical biology*, 99(4):629–644, December 1982.

[7] Benjamin Breiten, Matthew R Lockett, Woody Sherman, Shuji Fujita, Mohammad Al-Sayah, Heiko Lange, Carleen M Bowers, Annie Heroux, Goran Krilov, and George M Whitesides. Water networks contribute to enthalpy/entropy compensation in protein-ligand binding. *Journal of the American Chemical Society*, 135(41):15579–15584, October 2013.

[8] Phil Attard. Long-range attraction between hydrophobic surfaces. *The Journal of Physical Chemistry*, 93(17):6441–6444, August 1989.

[9] Jan Christer Eriksson, Stig Ljunggren, and Per M Claesson. A phenomenological theory of long-range hydrophobic attraction forces based on a square-gradient variational approach. *Journal of the Chemical Society, Faraday Transactions 2*, 85(3):163–176, 1989.

[10] Ka Lum, David Chandler, and John D Weeks. Hydrophobicity at Small and Large Length Scales. *The Journal of Physical Chemistry B*, 103(22):4570–4577, June 1999.

[11] Pieter Rein ten Wolde. Hydrophobic interactions: an overview. *Journal of Physics: Condensed Matter*, 14(40):9445–9460, September 2002.

[12] Huan-Xiang Zhou and Michael K Gilson. Theory of free energy and entropy in noncovalent binding. *Chemical Reviews*, 109(9):4092–4107, September 2009.

[13] Lin Li, Chuan Li, Zhe Zhang, and Emil Alexov. On the Dielectric "Constant" of Proteins: Smooth Dielectric Function for Macromolecular Modeling and Its Implementation in DelPhi. *Journal of Chemical Theory and Computation*, 9(4):2126–2136, April 2013.

[14] Holger Gohlke and Gerhard Klebe. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie (International ed. in English)*, 41(15):2644–2676, August 2002.

[15] George M Whitesides and Vijay M Krishnamurthy. Designing ligands to bind proteins. *Quarterly Reviews of Biophysics*, 38(04):385, July 2006.

[16] Stephen D Pickett and Michael J E Sternberg. Empirical scale of side-chain conformational entropy in protein folding. *Journal of Molecular Biology*, 231(3):825–839, June 1993.

[17] Yasmina S N Day, Cheryl L Baird, Rebecca L Rich, and David G Myszka. Direct comparison of binding equilibrium, thermodynamic, and rate constants determined by surface- and solution-based biophysical methods. *Protein Science*, 11(5):1017–1025, May 2002.

[18] Michael M Pierce, C S Raman, and Nall T Barry. Isothermal Titration Calorimetry of Protein–Protein Interactions. *Methods*, 19(2):213–221, October 1999.

[19] Robert Karlsson and Anders Fält. Experimental design for kinetic analysis of protein-protein interactions with surface plasmon resonance biosensors. *Journal of Immunological Methods*, 200(1-2):121–133, January 1997.

[20] Leonor Michaelis, Maud Leonora Menten, Kenneth A Johnson, and Roger S Goody. *The original Michaelis constant: translation of the 1913 Michaelis-Menten paper.*, volume 50. Biochemistry, October 2011.

[21] Yung-Chi Cheng and William H Prusoff. Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochemical Pharmacology*, 22(23):3099–3108, December 1973.

[22] Emil Fischer. Einfluss der Configuration auf die Wirkung der Enzyme . *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993, October 1894.

[23] Daniel E Koshland. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):98–104, February 1958.

[24] Robert L Baldwin. The nature of protein folding pathways: the classical versus the new view. *Journal of Biomolecular NMR*, 5(2):103–109, February 1995.

[25] Themis Lazaridis and Martin Karplus. "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science*, 278(5345):1928–1931, December 1997.

[26] Sandeep Kumar, Buyong Ma, Chung-Jung Tsai, Neeti Sinha, and Ruth Nussinov. Folding and binding cascades: dynamic landscapes and population shifts. *Protein Science*, 9(1):10–19, January 2000.

[27] Chung-Jung Tsai, Sandeep Kumar, Buyong Ma, and Ruth Nussinov. Folding funnels, binding funnels, and protein function. *Protein Science*, 8(6):1181–1190, May 1999.

[28] Chern-Sing Goh, Duncan Milburn, and Mark Gerstein. Conformational changes associated with protein–protein interactions. *Current Opinion in Structural Biology*, 14(1):104–109, February 2004.

[29] Simon J Teague. Implications of protein flexibility for drug discovery. *Nature Reviews Drug Discovery*, 2(7):527–541, July 2003.

[30] Yechun Xu, Jacques-Philippe Colletier, Martin Weik, Hualiang Jiang, John Moult, Israel Silman, and Joel L Sussman. Flexibility of aromatic residues in the active-site gorge of acetylcholinesterase: X-ray versus molecular dynamics. *Biophysical Journal*, 95(5):2500–2511, September 2008.

[31] Israel Silman and Joel L Sussman. Acetylcholinesterase: 'classical' and 'non-classical' functions and pharmacology. *Current Opinion in Pharmacology*, 5(3):293–302, June 2005.

[32] Terrone L Rosenberry. *Acetylcholinesterase*. Advances in Enzymology and Related Areas of Molecular Biology. John Wiley & Sons, Inc., Hoboken, NJ, USA, November 2006.

[33] Wei Wang, Wei Ye, Qingfen Yu, Cheng Jiang, Jian Zhang, Ray Luo, and Hai-Feng Chen. Conformational selection and induced fit in specific antibody and antigen recognition: SPE7 as a case study. *The Journal of Physical Chemistry B*, 117(17):4912–4923, May 2013.

[34] Michael B A Oldstone. Molecular mimicry and autoimmune disease. *Cell*, 50(6):819–820, September 1987.

[35] Todd F Kagawa, Jakki C Cooney, Heather M Baker, Sean McSweeney, Mengyao Liu, Siddeswar Gubba, James M Musser, and Edward N Baker. Crystal structure of the zymogen form of the group A Streptococcus virulence factor SpeB: an integrin-binding cysteine protease. *Proceedings of the National Academy of Sciences of the United States of America*, 97(5):2235–2240, October 1999.

[36] Christian B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, July 1973.

[37] Peter E Wright and Jane H Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293(2):321–331, October 1999.

[38] Megan Sickmeier, Justin A Hamilton, Tanguy LeGall, Vladimir Vacic, Marc S Cortese, Agnes Tantos, Beata Szabo, Peter Tompa, Jake Chen, Vladimir N Uversky, Zoran Obradovic, and A Keith Dunker. DisProt: the Database of Disordered Proteins. *Nucleic Acids Research*, 35(Database):D786–D793, January 2007.

[39] Vladimir N Uversky and A Keith Dunker. Understanding protein non-folding. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1804(6):1231–1264, June 2010.

[40] A Keith Dunker, Israel Silman, Vladimir N Uversky, and Joel L Sussman. Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology*, 18(6):756–764, December 2008.

[41] Lilia M Iakoucheva, Celeste J Brown, J David Lawson, Zoran Obradovic, and A Keith Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of Molecular Biology*, 323(3):573–584, October 2002.

[42] Andrew Campen, Ryan M Williams, Celeste J Brown, Jingwei Meng, Vladimir N Uversky, and A Keith Dunker. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein and Peptide Letters*, 15(9):956–963, 2008.

[43] Vladimir N Uversky, Christopher J Oldfield, and A Keith Dunker. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annual Review Biophysics*, 37:215–246, June 2008.

[44] Peter Tompa. Intrinsically unstructured proteins. *Trends in Biochemical Sciences*, 27(10):527–533, October 2002.

[45] Garegin A Papoian. Proteins with weakly funneled energy landscapes challenge the classical structure-function paradigm. *Proceedings of the National Academy of Sciences of the United States of America*, 105(38):14237–14238, September 2008.

[46] Steven J Metallo. Intrinsically disordered proteins are potential drug targets. *Current Opinion in Chemical Biology*, 14(4):481–488, August 2010.

[47] H Jane Dyson and Peter E Wright. Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology*, 12(1):54–60, February 2002.

[48] Peter Tompa and Monika Fuxreiter. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends in Biochemical Sciences*, 33(1):2–8, January 2008.

[49] Bálint Mészáros, István Simon, and Zsuzsanna Dosztányi. The expanding view of protein-protein interactions: complexes involving intrinsically disordered proteins. *Physical Biology*, 8(3):035003, June 2011.

[50] Vladimir N Uversky. Multitude of binding modes attainable by intrinsically disordered proteins: a portrait gallery of disorder-based complexes. *Chemical Society Reviews*, 40(3):1623–1634, March 2011.

[51] Patrick Chène. Inhibiting the p53-MDM2 interaction: an important target for cancer therapy. *Nature Reviews Cancer*, 3(2):102–109, February 2003.

[52] Sanjeev Shangary and Shaomeng Wang. Small-molecule inhibitors of the MDM2-p53 protein-protein interaction to reactivate p53 function: a novel

approach for cancer therapy. *Annual Review of Pharmacology and Toxicology*, 49:223–241, February 2009.

[53] Kareem Khoury, Grzegorz M Popowicz, Tad A Holak, and Alexander Dömling. The p53-MDM2/MDMX axis - A chemotype perspective. *MedChemComm*, 2:246–260, March 2011.

[54] Xiaoying Yin, Christine Giap, John S Lazo, and Edward V Prochownik. Low molecular weight inhibitors of Myc-Max interaction and function. *Oncogene*, 22(40):6151–6159, September 2003.

[55] Marcus Pickhardt, Zuzana Gazova, Martin von Bergen, Inna Khlistunova, Yipeng Wang, Antje Hascher, Eva-Maria Mandelkow, Jacek Biernat, and Eckhard Mandelkow. Anthraquinones inhibit tau aggregation and dissolve Alzheimer's paired helical filaments in vitro and in cells. *Journal of Biological Chemistry*, 280(5):3628–3635, February 2005.

[56] Tomer Cohen, Anat Frydman-Marom, Meirav Rechter, and Ehud Gazit. Inhibition of amyloid fibril formation and cytotoxicity by hydroxyindole derivatives. *Biochemistry*, 45(15):4727–4735, April 2006.

[57] Ariele Viacava Follis, Dalia I Hammoudeh, Huabo Wang, Edward V Prochownik, and Steven J Metallo. Structural Rationale for the Coupled Binding and Unfolding of the c-Myc Oncoprotein by Small Molecules. *Chemistry & Biology*, 15(11):1149–1155, November 2008.

[58] Réjean Lebel, Francois-Olivier McDuff, Pierre Lavigne, and Michel Grandbois. Direct visualization of the binding of c-Myc/Max heterodimeric b-HLH-LZ to E-box sequences on the hTERT promoter. *Biochemistry*, 46(36):10279–10286, August 2007.

[59] Dalia I Hammoudeh, Ariele Viacava Follis, Edward V Prochownik, and Steven J Metallo. Multiple independent binding sites for small-molecule inhibitors on the oncoprotein c-Myc. *Journal of the American Chemical Society*, 131(21):7390–7401, June 2009.

[60] Ariele Viacava Follis, Dalia I Hammoudeh, Andrew T Daab, and Steven J Metallo. Small-molecule perturbation of competing interactions between c-Myc and Max. *Bioorganic & Medicinal Chemistry Letters*, 19(3):807–810, February 2009.

[61] Julien Michel and Rémi Cuchillo. The impact of small molecule binding on the energy landscape of the intrinsically disordered protein C-myc. *PloS ONE*, 7(7):e41070, July 2012.

[62] Henryk M Wisniewski and Robert D Terry. Reexamination of the pathogenesis of the senile plaque. *Progress in Neuropathology*, 2:1–26, May 1973.

[63] Colin L Masters, Gerd Multhaup, Gail Simms, Jutta Pottgiesser, Ralph N Martins, and Konrad Beyreuther. Neuronal origin of a cerebral amyloid: neurofibrillary tangles of Alzheimer's disease contain the same protein as the amyloid of plaque cores and blood vessels. *The EMBO Journal*, 4(11):2757, November 1985.

[64] Prashant R Bharadwaj, Ashok K Dubey, Colin L Masters, Ralph N Martins, and Ian G Macreadie. AB aggregation and possible implications in Alzheimer's disease pathogenesis. *Journal of Cellular and Molecular Medicine*, 13(3):412–421, March 2009.

[65] Christina Priller, Thomas Bauer, Gerda Mitteregger, Bjarne Krebs, Hans A Kretzschmar, and Jochen Herms. Synapse formation and function is modulated by the amyloid precursor protein. *The Journal of Neuroscience*, 26(27):7212–7221, July 2006.

[66] Muray Coles, Wendy Bicknell, Andrew A Watson, David P Fairlie, and David J Craik. Solution structure of amyloid beta-peptide(1-40) in a water-micelle environment. Is the membrane-spanning domain where we think it is? *Biochemistry*, 37(31):11064–11077, August 1998.

[67] Michael D Carter, Gail Simms, and Donald F Weaver. The Development of New Therapeutics for Alzheimer's Disease. *Clinical Pharmacology & Therapeutics*, 88(4):475–486, September 2010.

[68] JoAnne McLaurin, Meredith E Kierstead, Mary E Brown, Cheryl A Hawkes, Mark H L Lambermon, Amie L Phinney, Audrey A Darabie, Julian E Cousins, Janet E French, Melissa F Lan, Fucheng Chen, Sydney S N Wong, Howard T J Mount, Paul E Fraser, David Westaway, and Peter St George-Hyslop. Cyclohexanehexol inhibitors of AB aggregation prevent and reverse Alzheimer phenotype in a mouse model. *Nature Medecine*, 12:801–808, June 2006.

[69] JoAnne McLaurin, Rivka Golomb, Anna Jurewicz, Jack P Antel, and Paul E Fraser. Inositol Stereoisomers Stabilize an Oligomeric Aggregate of Alzheimer

Amyloid beta Peptide and Inhibit Abeta -induced Toxicity. *Journal of Biological Chemistry*, 275(24):18495–18502, April 2000.

[70] Minna Groenning. Binding mode of Thioflavin T and other molecular probes in the context of amyloid fibrils—current status. *Journal of Chemical Biology*, 3(1):1–18, August 2009.

[71] Marino Convertino, Andreas Vitalis, and Amedeo Caflisch. Disordered binding of small molecules to AB (12–28). *Journal of Biological Chemistry*, 286(48):41578–41588, October 2011.

[72] Mika Ito, Jan Johansson, Roger Strömberg, and Lennart Nilsson. Effects of ligands on unfolding of the amyloid B-peptide central helix: mechanistic insights from molecular dynamics simulations. *PloS ONE*, 7(1):e30510, 2012.

[73] Grace Li, Sarah Rauscher, Stéphanie Baud, and Régis Pomès. Binding of inositol stereoisomers to model amyloidogenic peptides. *The Journal of Physical Chemistry B*, 116(3):1111–1119, January 2012.

[74] Maria Grazia Spillantini, Marie Luise Schmidt, Virginia M Y Lee, John Q Trojanowski, Ross Jakes, and Michel Goedert. |[alpha]|-Synuclein in Lewy bodies. *Nature*, 388(6645):839–840, August 1997.

[75] Minami Baba, Shigeo Nakajo, Pang-Hsien Tu, Taisuke Tomita, Kazuyasu Nakaya, Virginia M Y Lee, John Q Trojanowski, and Takeshi Iwatsubo. Aggregation of alpha-synuclein in Lewy bodies of sporadic Parkinson's disease and dementia with Lewy bodies. *The American Journal of Pathology*, 152(4):879, April 1998.

[76] K A Conway, J C Rochet, R M Bieganski, and P T Lansbury. Kinetic stabilization of the alpha-synuclein protofibril by a dopamine-alpha-synuclein adduct. *Science*, 294(5545):1346–1349, November 2001.

[77] Erin H Norris, Benoit I Giasson, Roberto Hodara, Shaohua Xu, John Q Trojanowski, Harry Ischiropoulos, and Virginia M Y Lee. Reversible inhibition of alpha-synuclein fibrillization by dopaminochrome-mediated conformational alterations. *The Journal of Biological Chemistry*, 280(22):21212–21219, June 2005.

[78] Fernando E Herrera, Alessandra Chesi, Katerina E Paleologou, Adrian Schmid, Adriana Munoz, Michele Vendruscolo, Stefano Gustincich, Hilal A Lashuel, and Paolo Carloni. Inhibition of alpha-synuclein fibrillization by

dopamine is mediated by interactions with five C-terminal residues and with E83 in the NAC region. *PloS ONE*, 3(10):e3394, 2008.

[79] Christofer Lendel, Carlos W Bertoncini, Nunilo Cremades, Christopher A Waudby, Michele Vendruscolo, Christopher M Dobson, Dale Schenk, John Christodoulou, and Gergely Toth. On the mechanism of nonspecific inhibitors of protein aggregation: dissecting the interactions of alpha-synuclein with Congo red and lacmoid. *Biochemistry*, 48(35):8322–8334, September 2009.

[80] Jie Li, Min Zhu, Amy B Manning-Bog, Donato A Di Monte, and Anthony L Fink. Dopamine and L-dopa disaggregate amyloid fibrils: implications for Parkinson's and Alzheimer's disease. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 18(9):962–964, June 2004.

[81] Huabo Wang, Dalia I Hammoudeh, Ariele Viacava Follis, Brian E Reese, John S Lazo, Steven J Metallo, and Edward V Prochownik. Improved low molecular weight Myc-Max inhibitors. *Molecular Cancer Therapeutics*, 6(9):2399–2408, September 2007.

[82] Charles Harrington, Janet E Rickard, David Horsley, Kathleen A Harrington, Kathleen P Hindley, Gernot Riedel, Franz Theuring, Kwang Meng Seng, and Claude M Wischik. Methylthioninium chloride (MTC) acts as a Tau aggregation inhibitor (TAI) in a cellular model and reverses Tau pathology in transgenic mouse models of Alzheimer's disease. *Alzheimer's & Dementia*, 4(4):T120–T121, July 2008.

[83] Bruno Bulic, Marcus Pickhardt, Eva-Maria Mandelkow, and Eckhard Mandelkow. Tau protein and tau aggregation inhibitors. *Neuropharmacology*, 59(4-5):276–289, September 2010.

3

An Example of IDPs : c-Myc

This chapter describes the impact of small molecule binding on the energy landscape of c-Myc

In the introduction chapter, it has been explained that a key requirement of structure-based drug design approaches is the availability of the three dimensional structure of a protein target. However, in solution, a protein cannot be considered as a rigid entity, rather it oscillates between different conformations with similar free energy.[1] Furthermore, a majority of proteins involved in diseases such as cardiovascular and neurodegenerative pathologies or cancers are known to be very flexible and the study of such proteins so called Intrinsically Disordered Proteins (IDPs) remains very challenging.[2] In this chapter, several computational methodologies were used to investigate the formation of 'hidden pockets' at the protein surface of the oncoprotein c-Myc and to study the impact of small molecule binding on the free energy landscape of this transcription factor.[3]

3.1 The oncoprotein c-Myc

c-Myc and the other proteins of the Myc family were among the first proto-oncogenes to have been identified.[4] These proteins are transcription factors able to activate the expression of several genes regulating many processes such as cell proliferation, cell differentiation or apoptosis. c-Myc, as other proteins belonging to this family, is organized into three different domains:[5]

1. A region for transcription activity in its N-terminal portion containing two highly conserved domain elements: MYC Box I (residues 45-63) and MYC Box II (residues 129-143) that are essential for the transactivation of the target genes.

2. A central region containing a nuclear localization site (residues 320-328), as well as two others MYC Box recently identified: MBIII (residues 188-199), which plays a role in cell transformation and MBIV (residues 295-315), involved in DNA binding, apoptosis, transformation and cell cycle arrest in G2.[6, 7]

3. A C-terminal domain, consisting of a basic region (residues 354-367) involved in recognition and binding to specific DNA sequences; a Helix-Loop-Helix domain (residues 368 to 407) and a Zip or Leucine Zipper motif (residues 413-434). This third region is illustrated in Figure 3.1A.

In its monomeric form, c-Myc is intrinsically disordered.[8] c-Myc has been shown to interact with a large number of other proteins. The specific interaction between c-Myc and the protein Max has been studied extensively because the c-Myc-Max heterodimer binds DNA and regulates gene expression.[9] It has been shown that overexpression of c-Myc is frequent in many cancers, and disruption of the c-Myc-Max interaction is a possible anticancer strategy.[10] Structurally diverse small molecules inhibiting the formation of this complex were discovered through a yeast two-hybrid screen.[11] Biophysical studies using fluorescence assays, Nuclear Magnetic Resonance (NMR) and Circular Dichroism (CD) measurements were performed to characterize protein-ligand interactions.[8, 12, 13] These studies suggest that the small molecules disrupt the c-Myc-Max interaction by stabilizing conformations in monomeric c-Myc that are incompatible with heterodimerization with Max. Three distinct binding sites, encompassing residues 366-375, 375-385 and 402-409, have been mapped on to the c-Myc bHLHZip domain.[12] Remarkably, the three distinct c-Myc-binding sites can be occupied simultaneously by different ligands. These results suggest that the c-Myc/small molecule interactions are fairly localized and can be predicted from primary sequence analysis. Indeed,

protein disorder prediction algorithms can locate approximately the small molecule-binding sites of c-Myc, which tend to be enriched in hydrophobic amino acids in comparison with the rest of the domain.[8] In addition, many of the small-molecule ligands can bind truncated c-Myc segments containing a single binding site with a binding affinity similar to that of the full c-Myc bHLHPZip domain. For instance, the small molecule 10058-F4 binds in a fluorescence polarization assay to c-Myc$_{353\text{-}437}$ with a Kd of 5.3±0.7 μM and to c-Myc$_{402\text{-}412}$ with a Kd of 13.3±1 μM.[8] Furthermore, similar chemical shift perturbations were observed for c-Myc$_{353\text{-}437}$ and c-Myc$_{402\text{-}412}$ upon binding 10058-F4. Therefore the small peptide c-Myc$_{402\text{-}412}$ appears to be a good model to study the interactions of 10058-F4 with full lenght c-Myc.

To detect and characterize hidden binding sites, MD simulations prove to be an attractive choice. In order to study the impact of small molecule binding on the energy landscape of the truncated peptide c-Myc$_{402\text{-}412}$, bias-exchange metadynamics simulations (BEMD) were performed in explicit solvent in absence and in presence of 10058-F4.[14]

3.2 Materials & Methods

3.2.1 Metadynamics Simulations

The protein and the ligand were built and prepared using the software Maestro.[15] The peptide termini were acetylated and amidated to be coherent with experimental data. All simulations were performed with the suite GROMACS

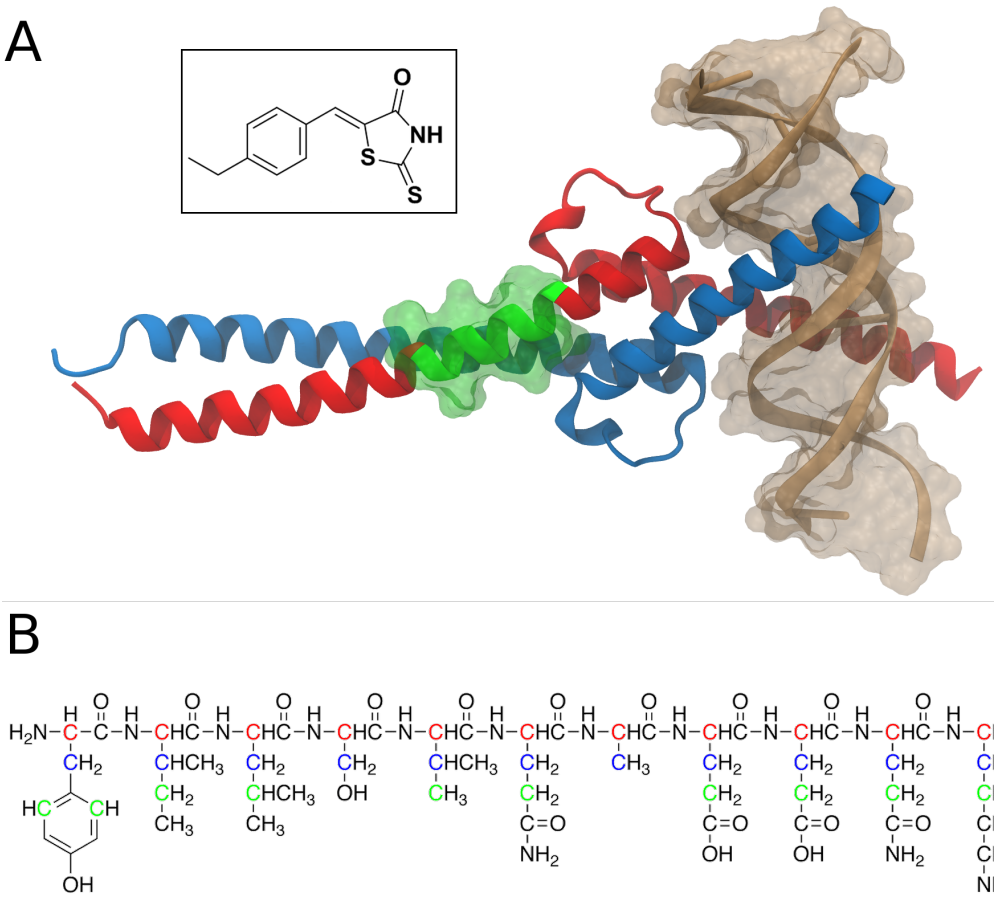

Figure 3.1: A) Structure of c-Myc (red) in complex with Max (blue) interacting with DNA (PDB code: 1NKP). The bHLHZip domain is formed by two helices separated by a small loop. According to the literature, the part of c-Myc targeted by the small molecule 10058-F4 (insert) is highlighted in green (sequence: YILSVQAEEQK). B) Chemical structure of the amino acid sequence used for the molecular dynamics simulations. The C_α, C_β and C_γ are respectively represented in red, blue and green.

4.5.5 compiled with the plugin PLUMED 1.3.[16, 17] The AMBER99SB* forcefield was selected for the small peptide c-Myc$_{402\text{-}412}$ while the GAFF force field was used for 10058-F4.[18, 19] The GAFF parameters for the ligand were obtained by using the python script ACPYPE in combination with the antechamber utility from the AMBER 11 software package.[20, 21] Atomic partial charges were assigned using the AM1-BCC method.[22, 23] Both systems apo c-Myc$_{402\text{-}412}$ and c-Myc$_{402\text{-}412}$/10058-F4 were solvated in a triclinic box with respectively 2843 and 3211 TIP3P water molecules and filled with enough counter ions to keep the system neutral. The minimal distance of the peptide to the boundary of the simulation box was at least 1.0 nm.[24] Temperature was controlled by a stochastic Berendsen thermostat and a coupling time of 0.1 ps. The default temperature for all simulations was 300 K. The pressure was controlled using a Parrinello-Rahman barostat at constant pressure 1 atm with a coupling time of 2.0 ps.[25] Long range electrostatic interactions were treated both with a short-range cut-off of 0.9 nm and the Particle-mesh Ewald method. A similar cut-off was used for the Lennard-Jones interactions. The neighbor list was updated every 10 integration steps. A long-range correction term was used for the energy and pressure.[26]After NPT equilibration, all production runs were performed for 120 ns in NVT conditions using a time step of 2.0 fs and LINCS constraints were applied to all covalent bonds.[27]

Preliminary runs were performed to optimize both the selection and the parametrization of CVs. The choice of those collective variables (CVs) were influenced by previously published BEMD studies to overcome possible energetic barriers between different peptide conformations.[28, 29] The parameters of the CVs (Gaussian height and width), which control the rate of convergence and accuracy of the free energy profiles were adjusted in preliminary runs in implicit solvent so as to obtain reasonably converged free energy profiles on a timescale of several dozen nanoseconds. Gaussian potentials of height 0.2 kJ.mol^{-1} were added every

2.0 ps. Collective variables and snapshots were saved every 2.0 ps and exchanges between replicas were attempted every 20.0 ps.

The simulations in presence (holo) and in absence (apo) of the ligand were performed with 8 and 9 replicas respectively. Each simulation was repeated twice using two different sets of starting conformations. These starting coordinates were obtained from preliminary runs and it was checked that they were structurally diverse and uncorrelated. Thus a total of 4 BEMD simulations were performed: two apo simulations (apoA and apoB) and two holo simulations (holoA and holoB) using three different CVs.[17]

The coordination number between the atoms i of a group G_1 and j in a group G_2 was calculated as

$$s = \sum_{i \in G_1} \sum_{j \in G_2} s_{ij} \tag{3.1}$$

with

$$s_{ij} = \begin{cases} 1 \text{ if } r_{ij} \leq 0 \\ \frac{1-\frac{r_{ij}}{r_0}^n}{1-\frac{r_{ij}}{r_0}^m} \text{if } r_{ij} > 0 \end{cases} \tag{3.2}$$

where $r_{ij} = \|r_i - r_j\| - d_0$. The parameters r_0, d_0, n and m were adjusted according to the type of interaction.

The minimum distance between two groups of atoms is measured as

$$s_{mindist} = \frac{\beta}{\log \sum_{ij} \exp(\beta / \|r_{ij}\|)} \tag{3.3}$$

where β=50.0

The number of hydrogen bonds is calculated as follows

$$s = \sum_{ij} \frac{1 - \frac{d_{ij}}{r_0}^n}{1 - \frac{d_{ij}}{r_0}^n} \tag{3.4}$$

where i is a hydrogen bond donor, j is a hydrogen bond acceptor and d_{ij} is the distance between between the atoms i and j.

The dihedral correlation is measured using the following equation:

$$s_{DC} = \sum_{i=2}^{N_D} \frac{1}{2} \left(1 + \cos\left(\phi_i - \phi_{i-1}\right)\right) \tag{3.5}$$

where N_D is the number of dihedrals in the CV.

The similarity of dihedral angles to a reference value is calculated as

$$s_{\alpha\beta} = \sum_{i=1}^{N_D} \frac{1}{2} \left(1 + \cos\left(\phi_i - \phi_i^{ref}\right)\right) \tag{3.6}$$

The parameters of each CV used to bias apo and holo simulations are given hereafter:

- **Apo simulations:** CV1: coordination number C_α atoms ($n = 8, m = 10, r_0 = 0.65$ nm, $d_0 = 0.0$ nm), width 0.7; CV2: coordination number C_γ atoms ($n = 8, m = 10, r_0 = 0.5$ nm, $d_0 = 0.0$ nm), width 0.5; CV3, similarity of backbone dihedral Ψ angle to α-helical region ($\phi_i = -1.31$), width 0.25; CV4, correlation of successive backbone dihedral angles; CV5: number of backbone - backbone hydrogen bonds ($r_0 = 0.25$ nm), width 0.25; CV6: number of sidechain - sidechain hydrogen bonds ($r_0 = 0.25$ nm), width

0.25; CV7: number of sidechain - backbone hydrogen bonds ($r_0 = 0.25$ nm), width 0.25;

- **Holo simulations:** Holo simulations: CV1: coordination number C_α atoms ($n = 8, m = 10, r_0 = 0.65$ nm, $d_0 = 0.0$ nm), width 0.7; CV2: coordination number C_γ atoms ($n = 8, m = 10, r_0 = 0.5$ nm, $d_0 = 0.0$ nm), width 0.5; CV3, similarity of backbone dihedral Ψ angle to α-helical region ($\phi_i = -1.31$), width 0.25; CV4, correlation of successive backbone dihedral angles; CV5: number of backbone - backbone hydrogen bonds ($r_0 = 0.25$ nm), width 0.25; CV6: number of sidechain - sidechain hydrogen bonds ($r_0 = 0.25$ nm), width 0.25; CV7: number of sidechain - backbone hydrogen bonds ($r_0 = 0.25$ nm), width 0.25; CV8: minimum distance ligand C1 atom to peptide C_α atoms. C1 is the aromatic carbon atom bonded to the methylene group of 10058-F4.

The gaussian accumulation allows the system to escape from a local minima and to gradually explore a broad range of values along each CV. This trend is more pronounced for CVs defined by counting interatomic contacts and eventually leads to the sampling of high energy configurations that cause hysteresis in the convergence of the free energy profiles for the biased replicas. However these limitations are significantly reduced by exchanging conformations between different runs according to the metropolis criterion. Moreover, high-energy configurations are almost never transferred to other replicas during replica exchange tests. To maintain a reasonable exchange rate between replicas and to focus conformational sampling in the regions of low free energy, half-harmonic potentials (walls) were added to penalize exploration of CV values below or above minimum/maximum values such that the computed free energy profiles are within approximately 10 k_BT from the global minimum. The position of the walls was chosen by performing

unrestrained preliminary BEMD runs.

- **Walls:** CV1: minimum 57, maximum 96; CV2: minimum 36, maximum 63; CV3: minimum 1, maximum 9; CV4, minimum 1.5, maximum 9.9; CV5, minimum 0.50, maximum 10.50; CV6, minimum 0.40, maximum 9.00; CV7 minimum 1.25, maximum 10.25; CV8 minimum 0.33, maximum 0.97.

With this setup the average exchange probability between biased replicas and neutral replicas was about 33% for both apo and holo simulations. According to the observed fluctuations in the values of the CVs over the duration of the BEMD simulations, all simulations have converged after 20 ns . Only the remaining 100 ns were considered for the analysis. The free energy profiles shown in Figure 3.2 and Figure 3.3 were taken as the negative of the averaged metadynamics biasing potential over the last 100 ns of each simulations. Computing equilibrium properties from low dimensional free energy projections is not an easy task. Indeed, when one wants to study a convoluted process such as the impact of ligand binding on the conformational sampling of a protein, each minimum in a low dimensional profile may correspond to several different structures. To overcome this limitation, the method of Marinelli et al. was used to reweigh snapshots from the biased simulations.[30] In this technique, the biased trajectories are first clustered in a N-dimensional CV space made of hypercubes forming a regular grid. The free energy of each bin is then estimated by a weighted histogram analysis procedure (WHAM) based on the number of snapshots and the value of the converged metadynamics bias potentials assigned to each bin (see Introduction for more details). As described by the authors, the accuracy of this approach is highly dependent of the bin properties. First, bins have to cover all the configuration space explored along the CV. Then, a large number of bins must be used and also be populated by a significant number of similar

conformations. After investigation using the VMD plugin METAGUI,[31] the best parameters identified for c-Myc$_{402\text{-}412}$ involved a 4-dimensional clustering using CV1, CV3, CV4, CV5 with a bin width of approximately 2σi, where σi is the Gaussian width of CVi.[31, 32] The choice of those 4 CVs were based on their poor correlation with each other, thus maximizing structural similarity of snapshots assigned to each bin. The bin width of 2σi is on the order of the resolution of the metadynamics free energy profiles. With this setup about 9000 bins were defined containing at least 5 snapshots. Lower dimensionality clustering produced bins that lumped together structurally dissimilar states, whereas higher dimensionality clustering yielded very few bins populated with more than five snapshots. Molecular observables were averaged between snapshots assigned to the same bin. Ensemble properties were then obtained by weighting the properties of each bin by its WHAM derived free energy. Concerning the ensemble properties of the neutral replica, they were simply computed by averaging the properties of each snapshot of the simulation. Beside the BEMD simulations, two classical MD simulations of c-Myc$_{402\text{-}412}$ were also performed (mdA and mdB). Similar simulations parameters to BEMD simulations were used but the time step that was set to 5 fs as virtual sites were used, and the simulations duration was 110 ns.[33] The first 10 ns were discarded to enable relaxation of the system. Thus, only the last 100 ns were considered for the analysis.

3.2.2 Simulations Analysis

In order to evaluate the equilibrium ensembles of c-Myc$_{402\text{-}412}$, the software Camshift was used to predict several NMR chemical shifts (^{1}H, ^{1}H$_{\alpha}$, ^{13}C$_{\alpha}$ and

$^{13}C_\beta$).[34] Camshift predictions are based on a polynomial expansion of the interatomic distances of the protein conformation. Because the approach is not able to assess the chemical shifts for N and C terminal residues no results are shown for Tyr_{402} and Lys_{412}. DSSP, STRIDE and PROSS were used to assess the secondary structure preferences from the simulations while the webserver δ2D was used to predict the same properties from the measured chemical shift.[35–39] A contact matrix was built to determine the preferred intramolecular and intermolecular interactions of c-$Myc_{402\text{-}412}$ in the apo and holo ensembles. A cutoff of 3 Å was used to define a proton-proton contact, which is intermediate between distances compatible with strong/medium NOEs. Small variations in this cutoff ($\pm$0.5Å) did not affect significantly the results. The approach developed by Daura et al. was applied to highlight the main conformations of the apo and holo equilibrium ensembles and estimates their proportion.[40] This iterative method relies on a RMSD clustering. First, RMSD calculations were performed between all pairs of structures in a trajectory. Then, for each snapshot, the number of structures that have a RMSD below a cutoff value are counted. The conformation with the highest number of similar structures is selected to define a cluster centre. This structure, along with all neighboring structures, is removed from the trajectory. Finally, the process is repeated until every structures are assigned to a cluster. In order to speed up the process while minimizing impact on the accuracy of the results, only snapshots from bins that were within $6k_BT$ from the bin of lowest free energy were selected. To estimate errors on the cluster populations, the ensembles from the two apo/holo simulations were combined using a RMSD cutoff of 3.5 Å. Different groups of atoms were retained to perform the RMSD calculations for the apo and holo ensembles.

For the apo simulations, all heavy atoms not involved in symmetry equivalent conformations (e.g Valine C_γ atom) were selected. Another selection is required

to consider the different possible binding modes of the ligand. The RMSD calculations were performed on the protein C_α and C_β atoms and non-symmetry equivalent ligand heavy atoms. Thus, the ligand coordinates were weighted by a factor of 3 in the RMSD calculations to in order to cluster together conformations that contained similar ligand coordinates.

3.3 Results

In order to characterize the structural ensembles of the peptide c-Myc$_{402\text{-}412}$ and the complex c-Myc$_{402\text{-}412}$/10058-F4 the bias-exchange variant of metadynamics was used.[14] Several biased simulations were run in parallel allowing a rapid exploration of the energy landscape of the system along a set of predefined collective variables. The technique is presented in detail in the Introduction chapter. Beside the biased simulations, an additional run without any bias, so called neutral replica, was able to exchange conformations with the other trajectories. According to the literature, the neutral replica has been found to produce an ensemble similar to the equilibrium ensemble of the system.[14, 30, 41, 42] Thus, BEMD has been shown to be an attractive tool to enhance the sampling of the folding free energy landscape of small proteins and protein/ligand complexes on timescales of a few dozen ns.[41, 43]

3.3.1 Conformational sampling of c-Myc$_{402\text{-}412}$

To ensure that the simulations have converged, one dimensional free energy profiles along the CVs used to enhance conformational sampling were computed. The reconstruction is obtained using the negative of the sum of the Gaussian biases added along the CV during the simulation.[44] These calculations were performed on two independent set of apo and holo simulations starting from structurally unrelated conformations. This was done to verify the reproducibility of our simulations.

The results are shown in Figures 3.2 & 3.3. In general the free energy profiles within 10 kJ.mol^{-1} of the global minimum are well reproduced (within ca. 1 k_BT or less) for most CVs between the two independent simulations. In the apo simulations, only the CV2 (coordination number C_γ atoms) shows a few discrepancies between the two simulations. The largest gap (5 kJ.mol^{-1}) was observed in the range of CV values of 40-50 contacts. Concerning the holo simulations, the biggest differences are more located in the regions of high free energy for CV2, CV3, CV4 and CV5 (Figure 3.3 B-E). In those regions, the divergence can reach 10 kJ.mol^{-1}. However, protein conformations with a high free energy contribute marginally to the equilibrium ensemble. Thus, the overall equilibrium properties back-computed from the biased simulations remain actually similar (e.g. Table 3.1 and Figure 3.5). As suggested by the Figure 3.4 A, the visual inspection of the apo neutral replica ensemble has confirmed that the system can adopt a broad range of conformations from compact to fully extended presenting occasionally secondary structure elements. The BEMD neutral replica ensembles were compared to two 100 ns unbiased MD simulation performed using

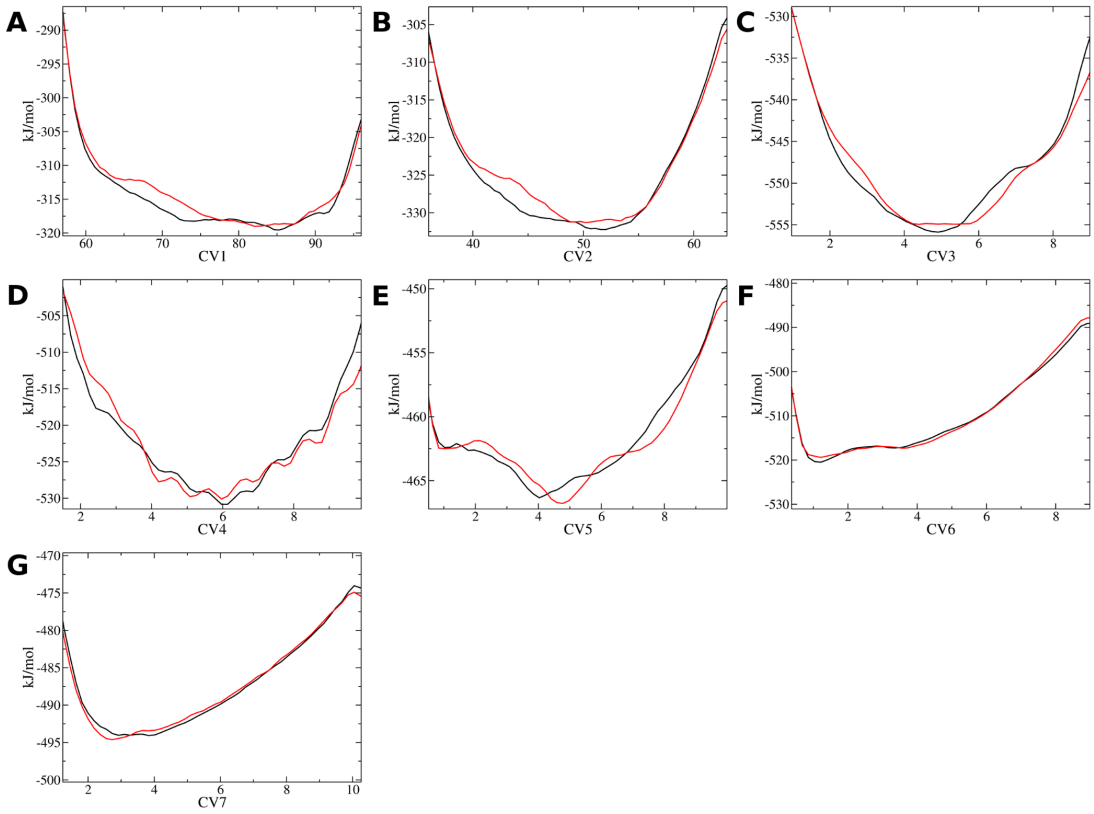

Figure 3.2: Free energy profiles for the c-Myc$_{402\text{-}412}$ apo simulations projected along several collective variables. Black: Simulation apoA, Red: Simulation apoB.A) CV1: coordination number C_α atoms. B)CV2: coordination number C_γ atoms. C) CV3: similarity of backbone dihedral Ψ angle to α-helical region. D) CV4: correlation of successive backbone dihedral angles. E) CV5: number of backbone - backbone hydrogen bonds. F) CV6: number of sidechain - sidechain hydrogen bonds. G) CV7: number of sidechain - backbone hydrogen bonds.

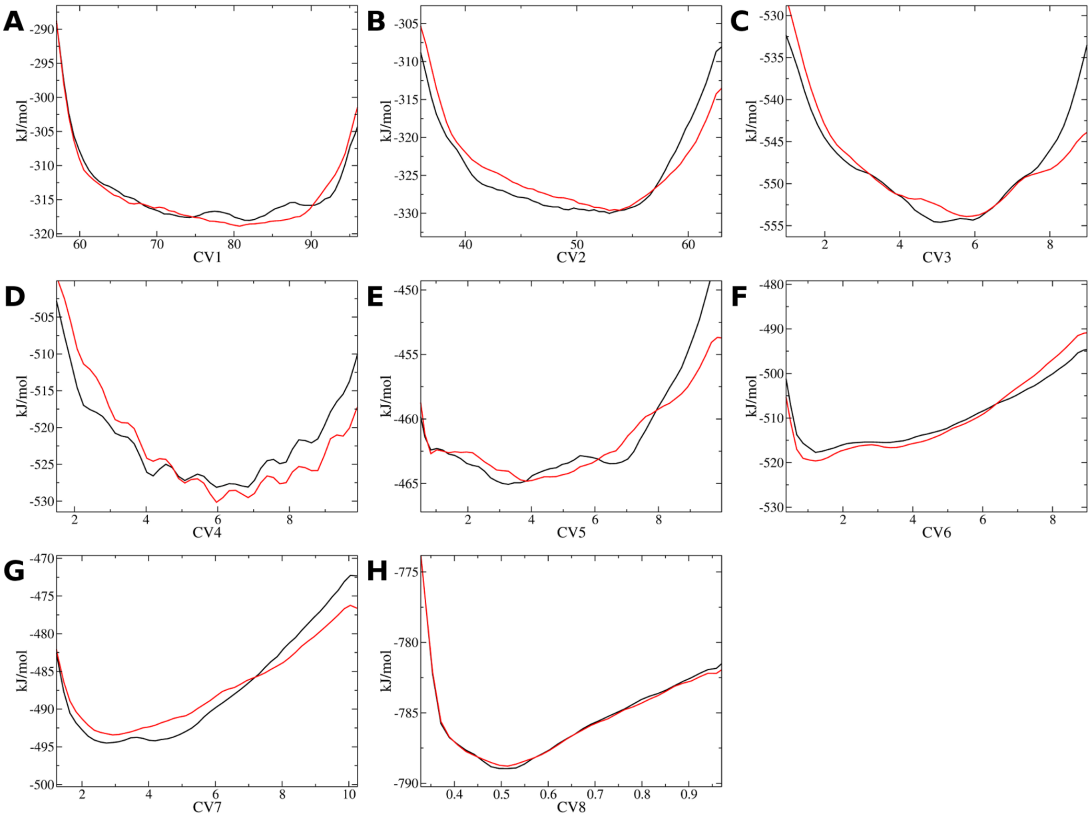

Figure 3.3: Free energy profiles for the c-Myc$_{402\text{-}412}$ holo simulations projected along several collective variables. Black: Simulation holoA, Red: Simulation holoB.A) CV1: coordination number C$_\alpha$ atoms. B) CV2: coordination number C$_\gamma$ atoms, width 0.5; CV3, similarity of backbone dihedral Ψ angle to α-helical region. C) CV4: correlation of successive backbone dihedral angles. D) CV5: number of backbone - backbone hydrogen bonds. E) CV6: number of sidechain - sidechain hydrogen bonds. F) CV7: number of sidechain - backbone hydrogen bonds. G) CV8: minimum distance ligand C1 atom to peptide Cα atoms.

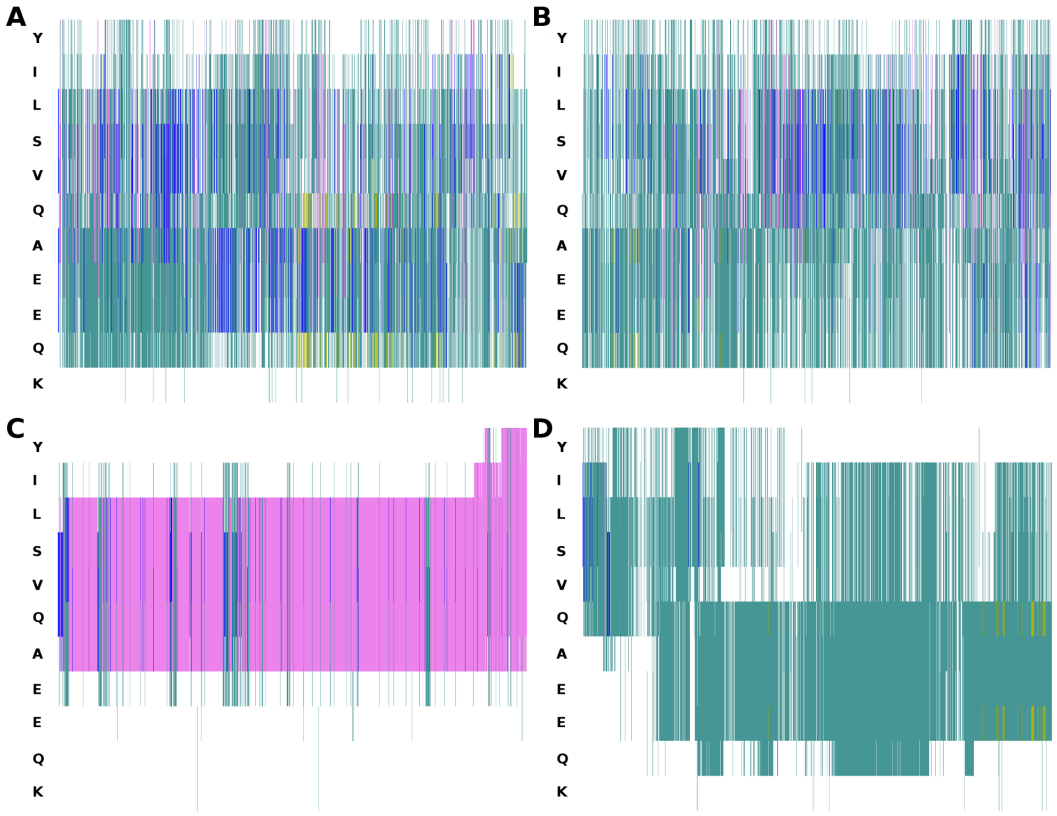

Figure 3.4: Secondary structure content of c-Myc$_{402\text{-}412}$. Residue secondary structure preferences colored according to the STRIDE code (white: coil, cyan: turn, blue: 3_{10} helix, purple: α-helix, maroon: bend, yellow: extended). A) and B) BEMD ensembles from the neutral replicas for simulations apoA and apoB. C) and D) Unbiased ensembles from MD simulations mdA and mdB.

the same potential energy function and system setup. The first MD simulation was initiated from an extended conformation which quickly forms a short α-helix from the amino acid Leu$_{404}$ to Ala$_{408}$ that is stable throughout the simulation (Figure 3.4 C). A very different conformational ensemble is observed for the second classical MD simulation (Figure 3.4 D). The system adopts mainly unstructured conformations. The lack of consistency between those two unbiased trajectories is a good illustration of the limitation of MD simulations to sample conformations from different local minimum compared to BEMD.

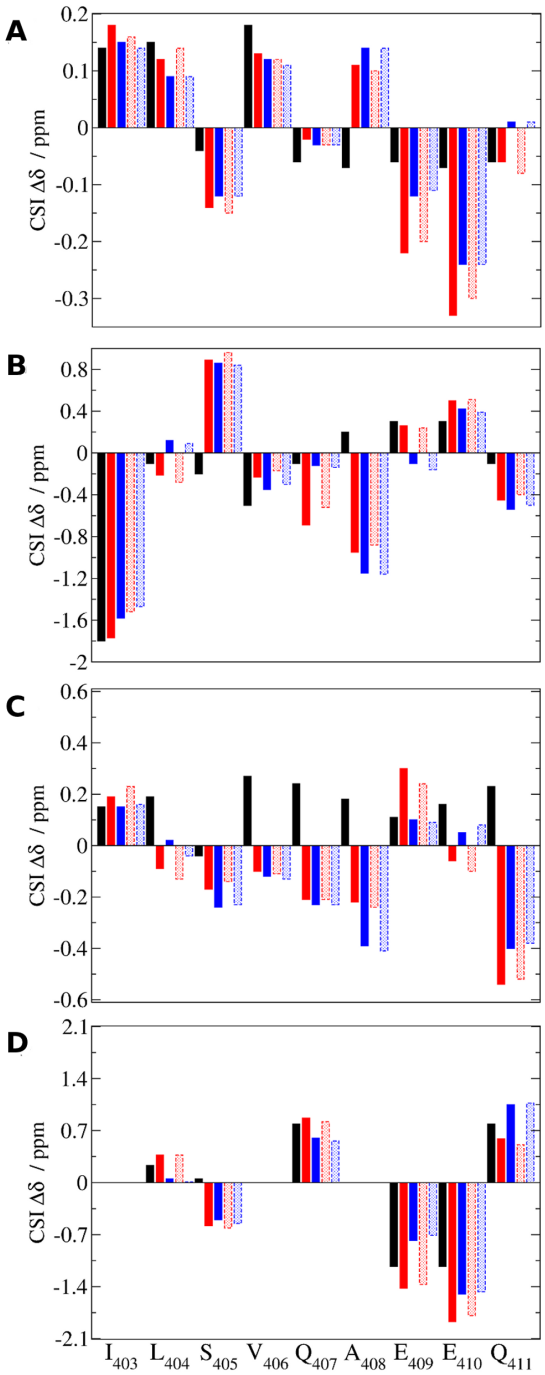

Figure 3.5: Comparison of computed and observed secondary chemical shifts for apo c-Myc$_{402\text{-}412}$. A) $^{1}H_{\alpha}$ chemical shifts. B) $^{13}C_{\alpha}$ chemical shifts. C) ^{1}H backbone amide chemical shifts. D) $^{13}C_{\beta}$ chemical shifts. Black: experimental data. Solid red and blue: predicted by reweighting the biased BEMD simulations apoA and apoB respectively. Dotted red and blue: predicted from the neutral replicas of the BEMD simulations apoA and apoB respectively. Not all experimental $^{13}C_{\beta}$ chemical shifts were reported. Camshift does not report chemical shifts for terminal residues.

	Helix		Sheet		Polyproline
BEMD run1[A]	13.9	10.0	0.7	0.9	12.6
BEMD run2[A]	10.9	9.4	0.6	0.2	12.1
BEMD run1 neutral[A]	14.2	11.9	0.7	1.0	11.6
BEMD run1 neutral[A]	10.5	9.2	0.5	0.2	11.7
MD run1[A]	34.3	41.5	0.0	0.0	5.5
MD run1[A]	0.1	0.1	0.5	0.5	13.5
Exp[B]	4.0		3.0		13.0

Table 3.1: Percentage of secondary structure content of apo c-Myc$_{402\text{-}412}$.[A]helix and sheet content were computed using the software DSSP and STRIDE respectively. A helix was defined as G + H + I according to the DSSP code. The polyproline II content was calculated with the software PROSS.[B]Structural features of the experimental data were estimated from the chemical shifts using the webserver δ2d.

To evaluate the accuracy of the equilibrium ensemble generated from the simulations, snapshots collected during the apo MD and BEMD simulations of c-Myc$_{402\text{-}412}$ were used to back-compute NMR chemical shifts using the software Camshift.[34] Subsequently, ^{1}H and ^{13}C secondary chemical shifts for H_α protons, backbone amide protons, C_α and C_β carbons were compared with experimental data (Figure 3.5).[13] Unfortunately, comparison of computed and measured chemical shifts for the c-Myc$_{402\text{-}412}$/10058-F4 complex is not possible owing to the lack of parameters in Camshift to describe the ligand. The secondary chemical shifts generated from the two BEMD ensembles show a good correlation with experimental values. Furthermore, only very small differences are observed between the chemical shifts obtained by averaging over snapshots from the neutral replicas or by reweighting snapshots from the biased simulations. By contrast, greater variability and inconsistency is observed between the chemical shifts computed from the two unbiased MD simulations (Figure 3.6). The mean-unsigned errors for the H_α, H, C_α and C_β chemical shifts computed from the two reweighted BEMD simulations are: 0.09/0.08, 0.43/0.44, 0.32/0.28 and 0.35/0.32 ppm respectively. Similar values were observed for the mean-unsigned errors

computed for the neutral replica ensembles: 0.09/0.08, 0.45/0.46, 0.32/0.29 and 0.34/0.35 ppm respectively. By comparison the mean-unsigned errors computed from the two MD simulations are: 0.15/0.13, 1.25/0.80, 0.50/0.42, 1.01/0.86 ppm respectively. As shown previously, the back-computed secondary chemical shifts suggest also that the protocol using BEMD produced more accurate and consistent equilibrium ensembles between independent runs. A last analysis was performed comparing the secondary structure content generated from our different equilibrium ensembles with experimental data (Table 3.1). The webserver δ2d was used to estimate the percentage of helix and sheet from the experimental chemical shifts.[39] The polyproline II content was calculated with the software PROSS.[38] The overall secondary structure content of the BEMD and MD ensembles was calculated using the softwares DSSP and STRIDE.[35–37] As shown in Table 3.1, both MD and BEMD are quite insensitive to the methodology used to predict the secondary structure. In general, the polyproline II, helix and sheet content of the BEMD simulations computed from the reweighted and the neutral replica ensembles was very similar and consistent. However, compared to experimental results, the proportion of helix was globally overestimated while the sheet content was underestimated. This systematic error could have been driven by the force field selected for the simulations. The most significant differences are seen in the MD simulations. In the first run helical conformations are mainly predicted. By contrast, only the polyproline II content matches experimental data. Given that both BEMD and MD simulations have been performed under similar conditions, the differences observed for the MD simulations could be explained by larger sampling errors. Although it is likely that optimized force fields could decrease further discrepancies with experiment, the computed BEMD ensemble is overall in reasonable agreement with the available experimental data for this system.

Along the different analysis completed on the BEMD simulations of c-

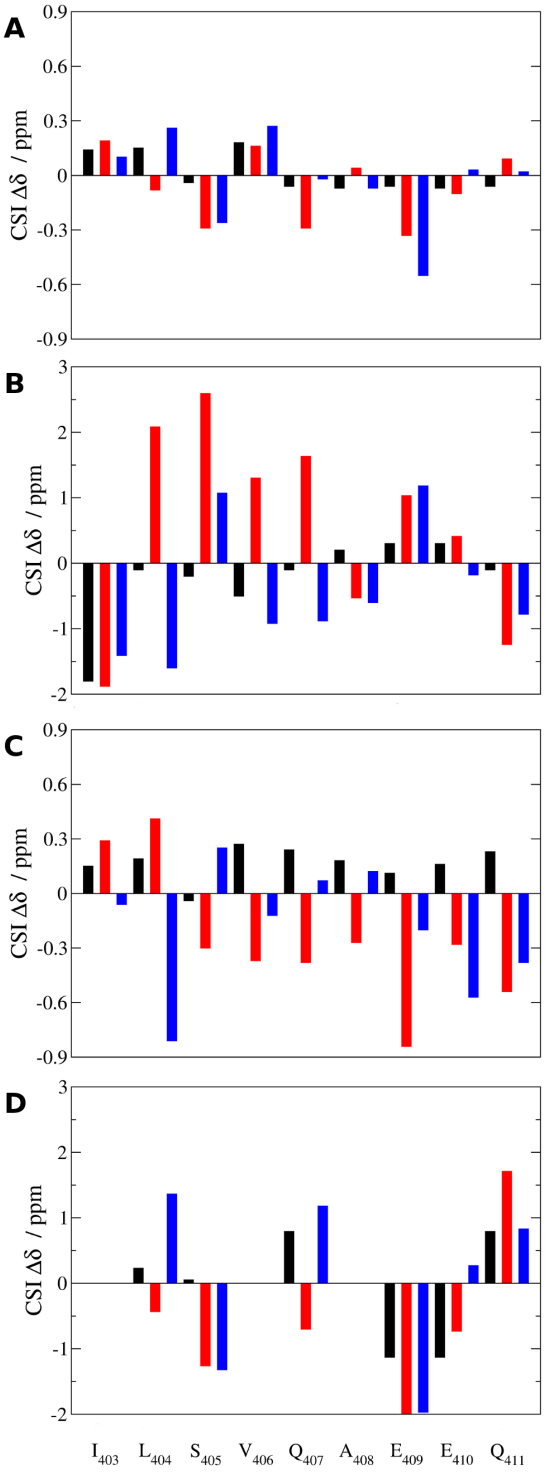

Figure 3.6: Comparison of computed and observed secondary chemical shifts for apo c-Myc$_{402\text{-}412}$ for the amino acids 403 to 411. A) $^{1}H_{\alpha}$ chemical shifts. B) $^{13}C_{\alpha}$ chemical shifts. C) ^{1}H backbone amide chemical shifts. D) $^{13}C_{\beta}$ chemical shifts. Black: experimental data. Red: predicted from MD simulation mdA. Blue: predicted from MD simulation mdB.

$Myc_{402\text{-}412}$, the results generated from the neutral replica was always very similar to the properties predicted by reweighting the biased simulations. This observation, in agreement with other bias- exchange metadynamics studies, suggests that the neutral replica is a good approximation of the equilibrium ensemble. As shown in Figure 3.7, the global minimum of the one-dimensional free energy profiles of all CVs are well reproduced. By contrast, the regions of high free energy are systematically overrepresented in the neutral replica. However, because conformations present in these CV values contributes marginally to the equilibrium ensemble, this does not affect significantly the different equilibrium properties (Figure 3.5 and Table 3.1). Nevertheless this analysis suggests that the accuracy of the neutral replica ensemble decreases rapidly for conformations of higher free energy. Consequently, analyses in the rest of the chapter were performed on ensembles constructed by reweighting snapshots from the biased simulations.

3.3.2 The c-$Myc_{402\text{-}412}$ Apo Ensemble

Clustering of the apo equilibrium ensemble of c-$Myc_{402\text{-}412}$ reveals dozen of structurally distinct conformations from collapsed to extended. Such heterogeneous sampling was predictable considering the intrinsically disordered feature of this system. Figure 3.8 depicts representative conformations from the nine largest clusters calculated for the apo ensemble. Similar clusters were found in the two independent simulations but sometimes not equally populated. The main cluster (Figure 3.8 A) is a random coil structure stabilized by hydrophobic contacts between Tyr_{402}, Ile_{403} and Val_{406} and electrostatic interactions between Lys_{412} and Glu_{409}. Other partially collapsed conformations are represented (Figures 3.8 D & 3.8 E) as well as extended conformations are also observed (e.g Figure 5F

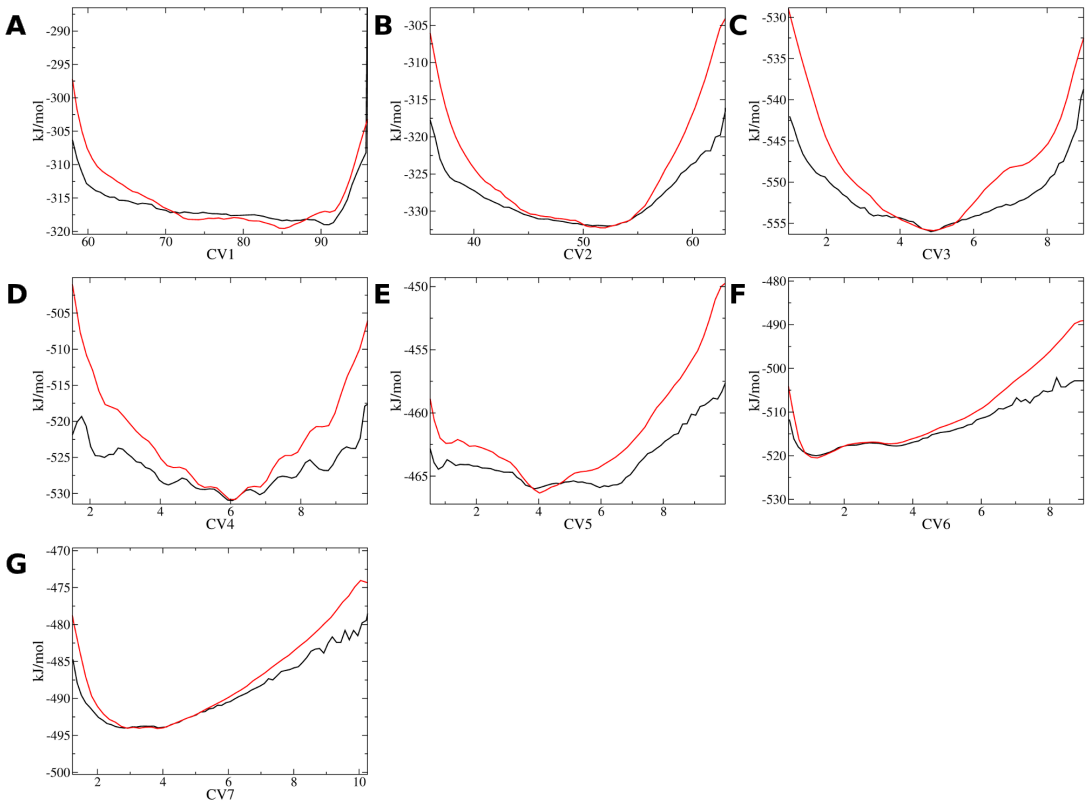

Figure 3.7: Comparison of free energy profiles of c-Myc$_{402\text{-}412}$ obtained from the neutral replica and the biased replicas. Black: Neutral replica, Red: Biased replica. Data generated using BEMD simulation apoA.

and 5I). Additionally, several clusters include conformations containing short α or 3_{10} helices (Figures 3.8 B & 3.8 G), that account for the overall computed helical content of c-Myc$_{402\text{-}412}$.

3.3.3 c-Myc$_{402\text{-}412}$ Remains Disordered upon Binding the Small Molecule 10058-F4

In order to assess the impact of the binding of 10058-F4 on the conformations of c-Myc$_{402\text{-}412}$, the average number of contacts between protons in 10058-F4 and different protein residues was computed for the apo and holo BEMD simulations.

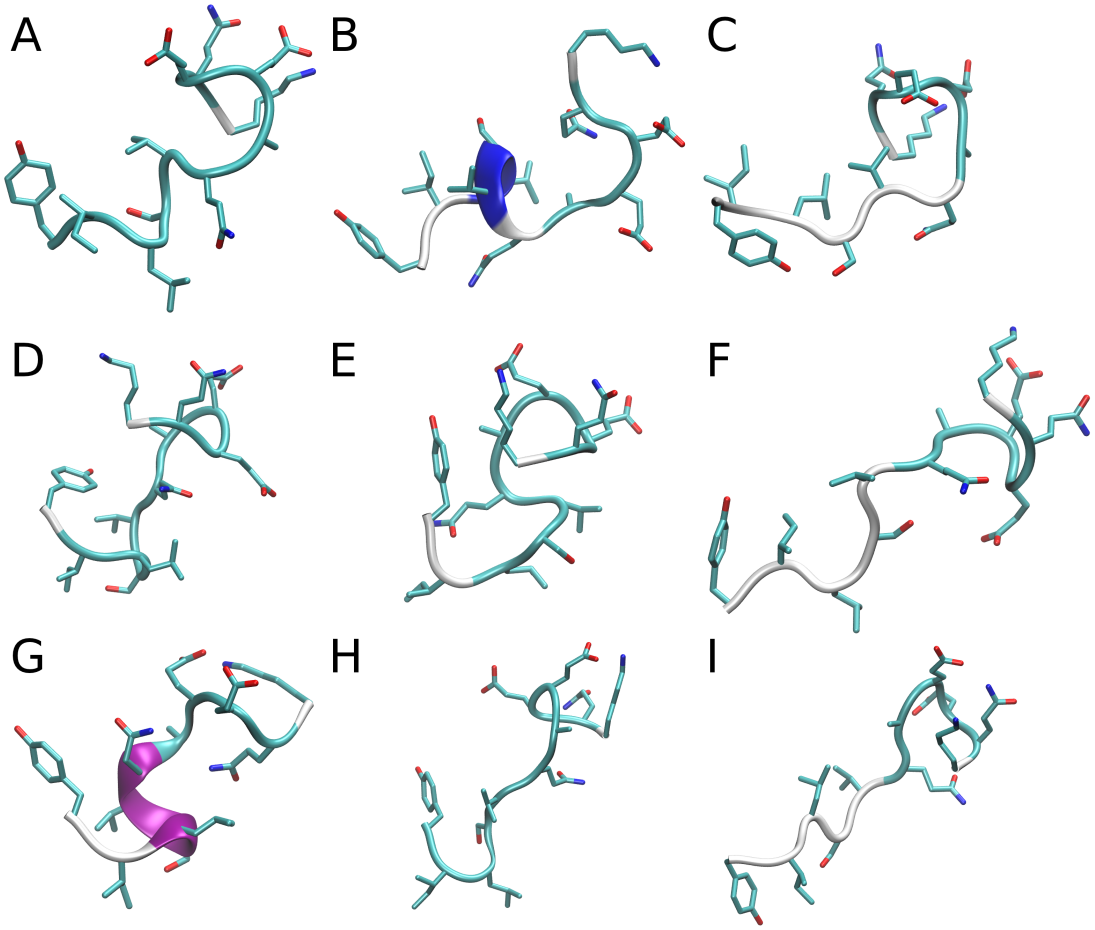

Figure 3.8: Representative conformations from the computed equilibrium ensemble for apo c-Myc$_{402\text{-}412}$. The conformations depicted are those closest to the center of the most populated clusters. The fractional cluster populations are: 0.101±0.018 (A), 0.075±0.034 (B), 0.060±0.040 (C), 0.059±0.027 (D), 0.055±0.016 (E), 0.043±0.004 (F), 0.030±0.017 (G), 0.021±0.009 (H), 0.021±0.003 (I).

Results using a cut-off of 3 Å are shown in Figure 3.9. The upper panel (Figure 3.9 A) suggests that 10058-F4 binds preferentially the N-terminal region of c-Myc$_{402\text{-}412}$ and more specifically Tyr$_{402}$. In the C-terminal region, only interactions with Lys$_{412}$ are visible. Given that 10058-F4 contains a moderately polar heterocycle and a hydrophobic ethylphenyl group, it is not surprising that intermolecular contacts occur preferentially with the N-terminal region as it is enriched in hydrophobic amino acids. Figure 3.9 B depicts the difference in average number of contacts between protein residues in the apo and holo simulations. A decrease of contacts between Tyr$_{402}$ and the neighboring amino acids is consistent with the previous results suggesting preferential interactions between 10058-F4 and the end of the N-terminal region. An increase in contacts between Lys$_{412}$ and the N-terminal part is clearly correlated with a decline number of interactions with the C-terminal domain. Those differences are explained by the observation that in the holo simulations, many conformations where c-Myc$_{402\text{-}412}$ is wrapping 10058-F4 are observed. Therefore, the terminal amino acids are more likely to be in contact with each other when the ligand is present. Additionally, the simulations suggest formation of a hydrophobic cluster between 10058-F4 and the side chains of Tyr$_{402}$, Ile$_{403}$, Leu$_{404}$, Val$_{406}$, which is in consistent with the experimental data published by Follis et al.[8] However, the holo equilibrium ensemble remains overall heterogeneous suggesting that 10058-F4 does not stabilize significantly c-Myc$_{402\text{-}412}$.

3.3.4 The Small Molecule 10058-F4 Binds Different c-Myc$_{402\text{-}412}$ Conformations

A visual inspection of the holo simulations reveals an important mobility of 10058-F4 all around c-Myc$_{402\text{-}412}$ involving a multitude of different binding modes.

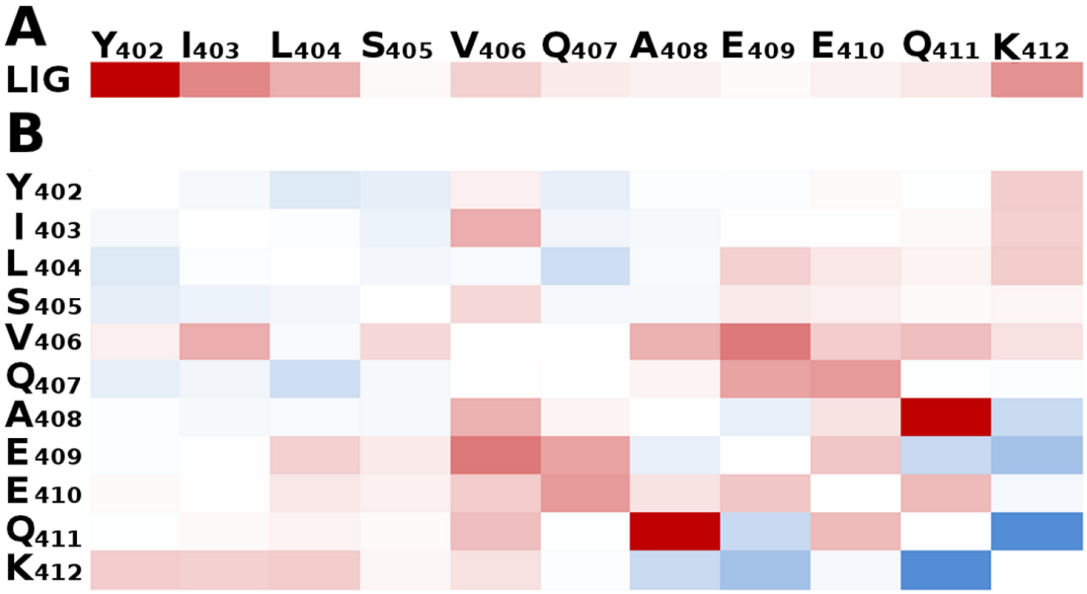

Figure 3.9: Average number of contacts between 1 and c-Myc$_{402\text{-}412}$. A) Average number of 1H contacts between different c-Myc$_{402\text{-}412}$ residues and 1. Color coded from white (no contacts) to red (high number of contacts). The extreme values of this color scale range from 0.02 to 1.08. B) Difference in the average number of 1H contacts between different c-Myc$_{402\text{-}412}$ residues in the holo and apo ensembles. Red/blue indicates an increased/decreased average number of contacts upon binding of 1. The extreme values of this color scale range from -1.22 to +0.67.

Consequently clustering analysis of the holo ensemble produces a large number of negligibly populated clusters. However, the most important clusters suggest that 10058-F4 interacts preferentially with specific c-Myc$_{402\text{-}412}$ conformations allowing to define the more likely binding modes (Figure 3.10).

The largest cluster (Figure 3.10 A) depicts stacking interactions between the phenyl rings of 10058-F4 and Tyr$_{402}$, as well as hydrophobic contacts between the ethylphenyl group of 10058-F4 and Leu$_{404}$. Ile$_{403}$ is involved in a small hydrophobic cluster with Leu$_{404}$ and Tyr$_{402}$. The conformation is also stabilized by several hydrogen-bonds between Gln$_{411}$ and the c-Myc$_{402\text{-}412}$ backbone. A different binding mode is depicted in Figure 3.10 B. The ethylphenyl group of 10058-F4 is stacked between the side-chains of Tyr$_{402}$ and Lys$_{412}$, while the thiazolidinone ring forms

hydrogen bonding interactions with the backbone of Leu_{404} and Gln_{407}. Other hydrophobic interactions between 10058-F4 and c-$Myc_{402\text{-}412}$ seem to stabilize the peptide in a helical conformation as shown in Figure 3.10 D & I. Only few contacts are observed for the four other clusters. Comparison of the computed holo c-$Myc_{402\text{-}412}$ conformations with the conformation of c-$Myc_{402\text{-}412}$ observed in the crystallographic structure of the c-Myc/Max dimer systematically indicates steric clashes with Max. Consequently, binding of 10058-F4 to c-Myc is not compatible with c-Myc/Max dimerization.

As discussed by Wang et al, the lack of well-defined structure of the c-$Myc_{402\text{-}412}$/10058-F4 complex could explain that just a few chemical modifications of 10058-F4 are able to improve significantly its binding affinity.[45] The largest populated holo cluster shows some important structural divergences with the c-$Myc_{402\text{-}412}$/10058-F4 complex derived using chemical-shift constraints and docking.[8] However, a single average structure generated from minimization of NMR derived restraints may not be representative of the multiple distinct conformations adopted by a disordered protein.[8] Therefore, molecular dynamics simulation is an attractive tool to generate structural ensembles for IDPs and guide the interpretation of NMR measurements.

3.3.5 c-$Myc_{402\text{-}412}$/10058-F4 Conformations are Partially Formed in the Apo Ensemble

In order to characterize the mechanisms of molecular recognition, the most representative apo and holo c-$Myc_{402\text{-}412}$ conformations (Figures 3.8 & 3.10) were compared to the computed apo and holo ensembles. The backbone root mean square deviation (RMSD) of the apo and holo structural ensembles according to

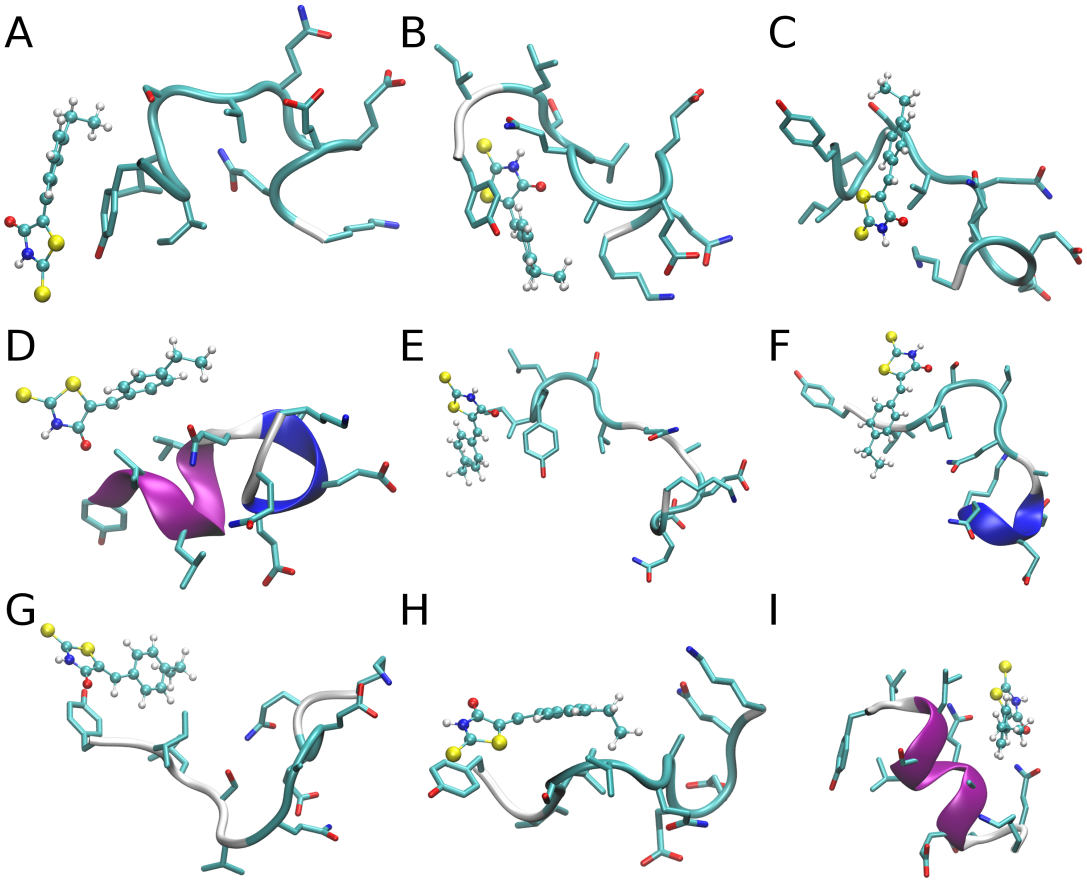

Figure 3.10: Representative conformations from the computed equilibrium ensemble for the c-Myc$_{402\text{-}412}$/10058-F4 complex. The conformations depicted are those closest to the cluster center. The fractional cluster populations are: 0.021±0.008 (A), 0.019±0.002 (B), 0.018±0.005 (C), 0.015±0.010 (D), 0.014±0.003 (E), 0.011±0.008 (F), 0.011±0.005 (G), 0.011±0.001 (H), 0.010±0.003 (I).

relevant apo and holo conformations is presented in Figure 3.11. The backbone RMSD cut-off to consider two protein structures as similar is never trivial to define, and is also highly dependent of the size of the system. For c-Myc$_{402\text{-}412}$, a backbone RMSD below 2.5 Å or less identifies roughly similar backbone conformations. Using this criterion, it was found that the c-Myc$_{402\text{-}412}$ apo ensemble contains backbone conformations that are structurally comparable to those seen more frequently in the holo ensemble (Figures 3.11 A-C), albeit with a lower probability. Figure 3.11 D highlights that frequently observed apo conformations are also present in the holo ensembles. To illustrate, Figure 3.11 also depicts an overlay of the conformation sampled from the apo (Figure 3.11A-C) or holo (Figure 3.11D) ensemble that has the lowest RMSD to the apo/holo conformations depicted in Figure 3.10A-C and Figure 3.8A. Even if these results suggest that there is significant structural overlap between the backbone of the apo and holo structures, side-chain rearrangements are necessary to allow apo c-Myc$_{402\text{-}412}$ conformations to accommodate 10058-F4.

3.4 Conclusion

Classical molecular dynamics and bias-exchange metadynamics simulations were performed on the small peptide c-Myc$_{402\text{-}412}$. The results add to the growing list of publications highlighting the usefulness of BEMD simulations to enhance conformational sampling of a protein.[30, 39, 41–43] Nevertheless, they also point out the difficulty of simulating the behaviour of IDPs using biomolecular force fields and a water model that are not always well adapted to describe flexible proteins with small energy differences between conformations interacting extensively with

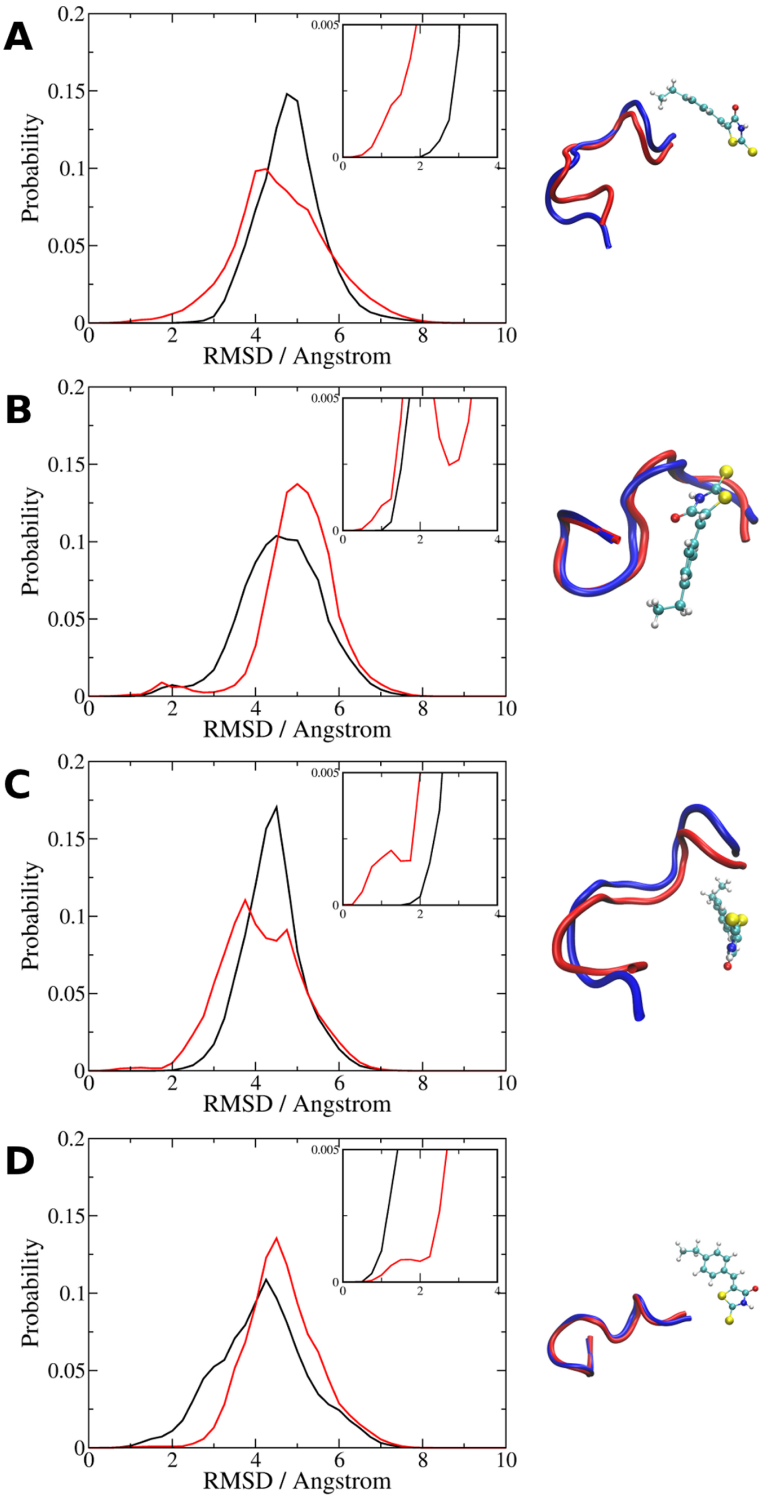

Figure 3.11: Comparison of selected holo and apo conformations to the apo and holo ensembles. A) Probability distribution of backbone RMSD of conformations from the apo (black curve) and holo (red curve) ensembles to: A) holo cluster center 3.10A, B) holo cluster center 3.10B, C) holo cluster center 3.10C, D) apo cluster center 3.8A. The inset shows the low-RMSD regions.Each panel also shows an overlay of the lowest RMSD apo or holo structure to cluster centers from panels A-D. For clarity only the peptide backbone (tube representation, apo conformations in blue, holo conformations in orange) and the ligand atoms (CPK) are shown.

the solvent. Therefore, comparing the equilibrium ensemble computed from the trajectories with experimental data such as NMR chemical shift is crucial to validate the simulations.[39] The systematic larger errors in predicted secondary structure content and chemical shifts for the MD simulation versus the BEMD simulations described along this chapter illustrate the consequences of a poor or insufficient conformational sampling for at least the regions of low free energy.[46]

The CVs used in this study have generated a broad range of conformations in both apo and holo equilibrium ensembles of c-Myc$_{402\text{-}412}$. As it has been suggested by Marinelli et al., that constructing a kinetic model of a system offers a better understanding of the free energy landscape.[30] Even if the kinetic properties are not directly available from the BEMD simulations, it is possible to project the BEMD trajectories on a space defined by the collective variables to build a kinetic model. However, the large heterogeneity of the structural ensemble of c-Myc$_{402\text{-}412}$ did not allow to clearly distinguish different kinetic basins in a low dimensional CV space. An interesting alternative would be to conduct extended unbiased MD simulations to reversibly simulate binding/unbinding in this system and analyze the computed trajectories using Markov State models.[47, 48] Furthermore, it has been shown that this kind of approach can achieve a direct estimation of dissociation constants. However in the present case, it may not be straightforward to define bound and unbound states for an IDP.[49]

The different simulations performed on c-Myc$_{402\text{-}412}$ did not allow the identification of a dominant binding mode with 10058-F4. Actually, 10058-F4 seems to interact with the peptide through a multitude of weak interactions with structurally diverse conformations. Those results are consistent with a recent study using mass spectroscopy suggesting that 10058-F4 may be not able to interact as strongly as it was described initially in the literature.[50] Indeed, the current

small molecule inhibitor of c-Myc$_{402\text{-}412}$ was reported to stabilize a broad range of conformations incompatible with dimerization with its partner Max, rather than order the peptide in a well-defined inactive form. Several other structurally different molecules were found to disrupt the c-Myc/Max complex, supporting the hypothesis that the large flexibility of IDPs promotes binding of diverse small molecules with distinct target conformations through weak interactions.[51] This observation is supported by other IDPs such as as CFTR/NBD1 or the cytoplasmic domain of the T-cell receptor ε chain/SIV nef protein complex that are known to remain partially disordered when in complex with a partner.[52, 53] According to the results presented, the molecular recognition of c-Myc$_{402\text{-}412}$ with 10058-F4 seems to fit both the conformational selection and induced fit models. Even if the most frequently observed holo conformations are visible, with a lower probability, in the apo equilibrium ensemble, a few side chain rearrangements are necessary to eliminate steric clashes with 10058-F4. This observation is in agreement with the extended conformational selection model where a conformational selection is combined with structural adjustments.[54]

The lack of specific interactions between 10058-F4 and c-Myc$_{353\text{-}437}$ makes the optimization of this compound difficult. The contact matrix in Figure 3.9 A, as well as the representative snapshots in Figure 3.10 suggest that 10058-F4 interacts preferentially with Tyr402. The amino acid sequence of the c-Myc bHLHZip domain is only composed of a unique Tyrosine. Moreover, as it is shown in Figure 3.12, the most hydrophobic part of the protein is located between the amino acids 401 and 406. This observation could explain the position of the binding site in this region. Interestingly, such hydrophobic clusters are not found in the bHLHZip domain of Max which seems to be generally more polar. This could explain why 10058-F4 has been reported to be unable to disrupt the Max/Max homodimer. However, many of the small molecules inhibitors of the c-Myc/Max identified

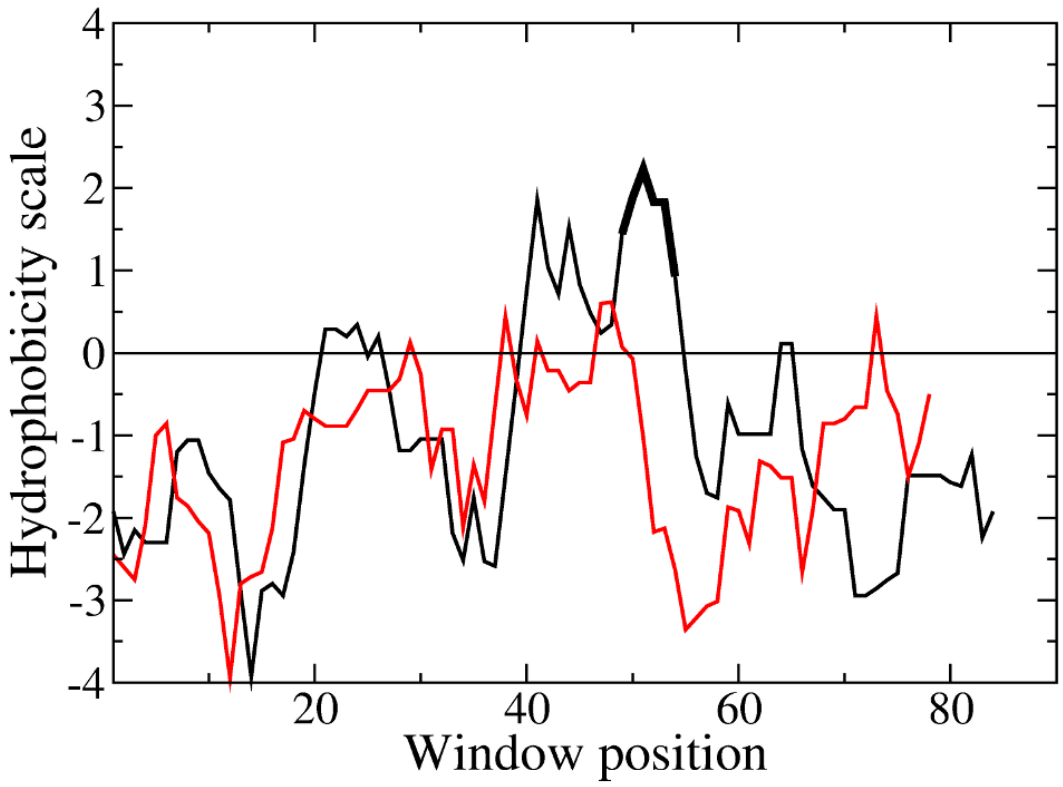

Figure 3.12: Hydrophobicity plot of the sequence of the c-Myc and Max bHLHZip domains. Black: c-Myc. Red: Max. Regions with a positive score are considered hydrophobic. The location of the c-Myc segment corresponding to amino acids 401 to 406 has been highlighted in bold. Plots generated using a Kyte-Doolittle hydrophobicity scale. To detect relatively short sequences of hydrophobic and aromatic sites that may interact favorably with small organic molecules the scale was modified so that Tyrosine has a hydrophobicity score equal to Phenylalanine and a window width of 3 was used. Plots produced using the sequences c-Myc$_{353\text{-}437}$ (84 amino acids) and Max$_{24\text{-}102}$ (78 amino acids).

from in vitro and cellular assays, including 10058-F4, are also able to disrupt other related protein-protein complex.[51, 55] In a larger study relying on yeast two hybrid assays performed on 32 protein complexes containing either HLH, HLHZip or bZip domains, 10058-F4 appeared to inhibit strongly c-Myc/Max, but also to a lesser extend Myod/E2-2, Mad1/Max, Mxi1/Max and Mad3/Max.[11] A notable feature of 10058-F4 is the presence of a benzylidene rhodamine. Actually, few other potent inhibitors of c-Myc are sharing this characteristic now known to frequently produce low-micromolar hits in a broad range of assays and against diverse targets.[56] To illustrate, a new benzylidene-rhodanine compound similar to 10058-F4 was recently found to bind the bZip region of another transcription factor δFosB.[57] Taken together, all the analysis performed to provide a better understanding of the c-Myc$_{402\text{-}412}$ /10058-F4 complex have highlighted only weak interactions consistent with a lack of specificity between the ligand and its target. Therefore, modifying 10058-F4 in order to enhance binding affinity towards c-Myc proves very challenging and necessarily involves increasing the number of specific interactions. Furthermore, simulations did not allow to clearly identify hidden pockets at the surface of the peptide c-Myc$_{402\text{-}412}$. This provides additional motivation to develop a novel molecular simulation methodology to detect binding sites that are not seen in the static picture of a protein structure revealed by experiments.

3.4 Bibliography

[1] Simon J Teague. Implications of protein flexibility for drug discovery. *Nature Reviews Drug Discovery*, 2(7):527–541, July 2003.

[2] A Keith Dunker, Israel Silman, Vladimir N Uversky, and Joel L Sussman. Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology*, 18(6):756–764, December 2008.

[3] Julien Michel and Rémi Cuchillo. The impact of small molecule binding on the energy landscape of the intrinsically disordered protein C-myc. *PloS ONE*, 7(7):e41070, July 2012.

[4] Gary Ramsay, Gerard I Evan, and J Michael Bishop. The protein encoded by the human proto-oncogene c-myc. *Proceedings of the National Academy of Sciences of the United States of America*, 81(24):7742–7746, December 1984.

[5] Linda M Facchini and Linda Z Penn. The molecular role of Myc in growth and transformation: recent discoveries lead to new insights. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 12(9):633–651, June 1998.

[6] Andreas Herbst, Michael T Hemann, Kathryn A Tworkowski, Simone E Salghetti, Scott W Lowe, and William P Tansey. A conserved element in Myc that negatively regulates its proapoptotic activity. *EMBO reports*, 6(2):177–183, February 2005.

[7] Victoria H Cowling, Sanjay Chandriani, Michael L Whitfield, and Michael D Cole. A conserved Myc protein domain, MBIV, regulates DNA binding, apoptosis, transformation, and G2 arrest. *Molecular and Cellular Biology*, 26(11):4226–4239, June 2006.

[8] Ariele Viacava Follis, Dalia I Hammoudeh, Huabo Wang, Edward V Prochownik, and Steven J Metallo. Structural Rationale for the Coupled Binding and Unfolding of the c-Myc Oncoprotein by Small Molecules. *Chemistry & Biology*, 15(11):1149–1155, November 2008.

[9] Réjean Lebel, Francois-Olivier McDuff, Pierre Lavigne, and Michel Grandbois. Direct visualization of the binding of c-Myc/Max heterodimeric b-HLH-LZ to E-box sequences on the hTERT promoter. *Biochemistry*, 46(36):10279–10286, August 2007.

[10] Lilia M Iakoucheva, Celeste J Brown, J David Lawson, Zoran Obradovic, and A Keith Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of Molecular Biology*, 323(3):573–584, October 2002.

[11] Xiaoying Yin, Christine Giap, John S Lazo, and Edward V Prochownik. Low molecular weight inhibitors of Myc-Max interaction and function. *Oncogene*, 22(40):6151–6159, September 2003.

[12] Dalia I Hammoudeh, Ariele Viacava Follis, Edward V Prochownik, and Steven J Metallo. Multiple independent binding sites for small-molecule inhibitors on the oncoprotein c-Myc. *Journal of the American Chemical Society*, 131(21):7390–7401, June 2009.

[13] Ariele Viacava Follis, Dalia I Hammoudeh, Andrew T Daab, and Steven J Metallo. Small-molecule perturbation of competing interactions between c-Myc and Max. *Bioorganic & Medicinal Chemistry Letters*, 19(3):807–810, February 2009.

[14] Stefano Piana and Alessandro Laio. A Bias-Exchange Approach to Protein Folding. *The Journal of Physical Chemistry B*, 111(17):4553–4559, May 2007.

[15] Maestro, version 9.7, Schrödinger, LLC, New York, NY, 2014.

[16] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, March 2008.

[17] Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provasi, Paolo Raiteri, Davide Donadio, Fabrizio Marinelli, Fabio Pietrucci, Ricardo A Broglia, and Michele Parrinello. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications*, 180(10):1961–1972, October 2009.

[18] Robert B Best and Gerhard Hummer. Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides. *The Journal of Physical Chemistry B*, 113(26):9004–9015, July 2009.

[19] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, July 2004.

[20] Alan W Sousa da Silva and Wim F Vranken. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC research notes*, 5:367, July 2012.

[21] D A Case, T A Darden, T E Cheatham, and C L Simmerling. Amber 11. 2010.

[22] Araz Jakalian, Bruce L Bush, David B Jack, and Christopher I Bayly. Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method. *Journal of Computational Chemistry*, 21(2):132–146, January 2000.

[23] Araz Jakalian, David B Jack, and Christopher I Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry*, 23(16):1623–1641, December 2002.

[24] Herman J C Berendsen, Johan P M Postma, Wilfred F van Gunsteren, Alfredo Di Nola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *Journal of Chemical Information and Modeling*, 81(8):3684–3690, June 1984.

[25] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182, December 1981.

[26] Michael R Shirts, David L Mobley, John D Chodera, and Vijay S Pande. Accurate and efficient corrections for missing dispersion interactions in molecular simulations. *Journal of Physical Chemistry Letters*, 111(45):13052–13063, November 2007.

[27] Berk Hess. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(1):116–122, January 2008.

[28] Giovanni Bussi, Francesco Luigi Gervasio, Alessandro Laio, and Michele Parrinello. Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics. *Journal of the American Chemical Society*, 128(41):13435–13441, October 2006.

[29] Stefano Piana, Alessandro Laio, Fabrizio Marinelli, Marleen Van Troys, David Bourry, Christophe Ampe, and José C Martins. Predicting the effect of a point mutation on a protein fold: the villin and advillin headpieces and their Pro62Ala mutants. *Journal of Molecular Biology*, 375(2):460–470, January 2008.

[30] Fabrizio Marinelli, Fabio Pietrucci, Alessandro Laio, and Stefano Piana. A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. *PLoS Computational Biology*, 5(8):e1000452, August 2009.

[31] Xevi Biarnés, Fabio Pietrucci, Fabrizio Marinelli, and Alessandro Laio. METAGUI. A VMD interface for analyzing metadynamics and molecular dynamics simulations. *Computer Physics Communications*, 183(1):203–211, January 2012.

[32] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–8– 27–8, February 1996.

[33] David van der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman J C Berendsen. GROMACS: fast, flexible, and free. *Journal of Computational Chemistry*, 26(16):1701–1718, December 2005.

[34] Kai J Kohlhoff, Paul Robustelli, Andrea Cavalli, Xavier Salvatella, and Michele Vendruscolo. Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *Journal of the American Chemical Society*, 131(39):13894–13895, October 2009.

[35] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983.

[36] Robbie P Joosten, Tim A H te Beek, Elmar Krieger, Maarten L Hekkelman, Rob W W Hooft, Reinhard Schneider, Chris Sander, and Gert Vriend. A series of PDB related databases for everyday needs. *Nucleic Acids Research*, 39(Database issue):D411–9, January 2011.

[37] Dmitrij Frishman and Patrick Argos. Knowledge-based protein secondary structure assignment. *Proteins: Structure*, 23(4):566–579, December 1995.

[38] Rajgopal Srinivasan and George D Rose. A physical basis for protein secondary structure. In *Proceedings of the National Academy of Sciences*, pages 14258–14263, October 1999.

[39] Carlo Camilloni, Alfonso De Simone, Wim F Vranken, and Michele Vendruscolo. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry*, 51(11):2224–2231, February 2012.

[40] Xavier Daura, Karl Gademann, Bernhard Jaun, Dieter seebach, Wilfred F van Gunsteren, and Alan E Mark. Peptide folding: when simulation meets experiment. *Angewandte Chemie International Edition*, 38(1-2):236–240, January 1999.

[41] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5):826–843, September 2011.

[42] Nevena Todorova, Fabrizio Marinelli, Stefano Piana, and Irene Yarovsky. Exploring the folding free energy landscape of insulin using bias exchange metadynamics. *Journal of Physical Chemistry Letters*, 113(11):3556–3564, March 2009.

[43] Vanessa Leone, Fabrizio Marinelli, Paolo Carloni, and Michele Parrinello. Targeting biomolecular flexibility with metadynamics. *Current Opinion in Structural Biology*, 20(2):148–154, April 2010.

[44] Alessandro Laio and Francesco L Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, 71(12):126601, November 2008.

[45] Huabo Wang, Dalia I Hammoudeh, Ariele Viacava Follis, Brian E Reese, John S Lazo, Steven J Metallo, and Edward V Prochownik. Improved low molecular weight Myc-Max inhibitors. *Molecular Cancer Therapeutics*, 6(9):2399–2408, September 2007.

[46] David L Mobley. Let's get honest about sampling. *Journal of Computer-Aided Molecular Design*, 26(1):93–95, January 2012.

[47] Vijay S Pande, Kyle Beauchamp, and Gregory R Bowman. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods*, 52(1):99–105, September 2010.

[48] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics*, 134(17):174105, May 2011.

[49] Daniel-Adriano Silva, Gregory R Bowman, Alejandro Sosa-Peinado, and Xuhui Huang. A role for both conformational selection and induced fit in ligand binding by the LAO protein. *PLoS Computational Biology*, 7(5):e1002054, May 2011.

[50] Sophie R Harvey, Massimilliano Porrini, Christiane Stachl, Derek MacMillan, Giovanna Zinzalla, and Perdita E Barran. Small-molecule inhibition of c-MYC: MAX leucine zipper formation is revealed by ion mobility mass spectrometry. *Journal of the American Chemical Society*, 134(47):19384–19392, October 2012.

[51] Steven J Metallo. Intrinsically disordered proteins are potential drug targets. *Current Opinion in Chemical Biology*, 14(4):481–488, August 2010.

[52] Jennifer M R Baker, Rhea P Hudson, Voula Kanelis, Wing-Yiu Choy, Patrick H Thibodeau, Philip J Thomas, and Julie D Forman-Kay. CFTR regulatory region interacts with NBD1 predominantly via multiple transient helices. *Nature structural & molecular biology*, 14(8):738–745, August 2007.

[53] Alexander B Sigalov, Walter M Kim, Maria Saline, and Lawrence J Stern. The intrinsically disordered cytoplasmic domain of the T cell receptor zeta chain binds to the nef protein of simian immunodeficiency virus without a disorder-to-order transition. *Biochemistry*, 47(49):12942–12944, December 2008.

[54] Peter Csermely, Robin Palotai, and Ruth Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in Biochemical Sciences*, 35(10):539–546, October 2010.

[55] Inga Müller, Karin Larsson, Anna Frenzel, Ganna Oliynyk, Hanna Zirath, Edward V Prochownik, Nicholas J Westwood, and Marie Arsenian Henriksson. Targeting of the MYCN protein with small molecule c-MYC inhibitors. *PloS ONE*, 9(5):e97285, 2014.

[56] Jonathan B Baell and Georgina A Holloway. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740, April 2010.

[57] Yun Wang, Teresa I Cesena, Yoko Ohnishi, Rebecca Burger-Caplan, Vivian Lam, Paul D Kirchhoff, Scott D Larsen, Martha J Larsen, Eric J Nestler, and Gabby Rudenko. Small molecule screening identifies regulators of the transcription factor deltaFosB. *ACS Chemical Neuroscience*, 3(7):546–556, July 2012.

4

JEDI SCORING FUNCTION

This chapter introduces JEDI, a novel methodology to assess the druggability 'on the fly' during a molecular dynamics simulation

The development of a new medicine is a long and expensive process subjected to high attrition rates.[1] Over the last decades, around 60% of drug discovery projects failed to identify viable leads able to modulate adequately the activity of a protein target.[2] Analyses of the sequenced human genome indicate that less than 50% of disease-involved genes code for druggable proteins.[3, 4] A protein target found to be nondruggable late in the drug discovery process is a significant waste of time and expense in the pharmaceutical industry. Accordingly, an early assessment of druggability offers the opportunity to focus efforts on tractable targets, thereby reducing the rate of failure.[5] The concept of druggability is ambiguous because it has been used in many different fields to describe, in a different context, the properties of genes, proteins and ligands. In the context of structure-based drug design, protein druggability is often related to the ability of a therapeutic target to bind a drug-like small molecule, leaving aside many important facets of the drug discovery and development process such as selectivity, toxicology or pharmacokinetics.[3] Since druggability is closely linked to the notion of binding site in this specific context, the terms 'bindability' or 'ligandability' have also been proposed as alternatives.[6, 7] This report focuses on the use of a computational approach for structure-based evaluation of protein druggability.

4.1 Introduction

The idea of relating binding site energetics to structural descriptors was explored as early as in 1985 with the Grid program of Goodford, and other related methods.[8–11] As interest in druggability developed in the last fifteen years, more recent efforts have focused on correlating directly structural descriptors

to druggability. An early effort was contributed by Hadjuk and coworkers.[12] NMR-based fragment screening was used to develop a mathematical model for druggability measurements whereby structural descriptors were correlated to NMR hit-rates. The methodology is based on the assumption that a druggable cavity tends to bind more fragments than a nondruggable pocket. A second approach, called MAP_{POD}, was published by Cheng et al. shortly after.[13] The authors proposed a scoring function to assess the maximal affinity between a small molecule and a binding site based on physicochemical and geometric features. This study also introduced a new category of proteins that are neither 'druggable' or 'nondruggable', but are instead 'difficult' to target with small molecules. The suggestion was that this category of proteins should be targeted with highly polar molecules administrated as pro-drugs. These early contributions have paved the way for a similar class of computational methods that aim to detect and evaluate potential binding sites at protein surfaces. The public dataset compiled for MAP_{POD} was used to parameterize Dscore, a druggability function coupled with the pocket detector SiteMap.[14, 15] Dscore is a simple linear combination of three descriptors reflecting the volume, enclosure and hydrophobicity of the binding site. Schmidtke et al. have recently developed a fast methodology based on a new publically accessible dataset.[16, 17] The approach features a logistic regression analysis to extract local and global hydrophobic descriptors of a protein pocket. One of the most recent structure-based approaches published in the field is called Drugpred.[18] Drugpred is based on the largest freely accessible non-redundant dataset and it appears to be less sensitive to binding site structural modifications that do not dramatically affect pocket properties.[18]

The above described methods were designed to assess druggability of a crystallographic protein structure. However, it is well known that sometimes a few local structural rearrangements around a protein binding site can profoundly

influence the affinity of a small molecule to its target.[19, 20] Accordingly, a second class of druggability prediction algorithms based on molecular dynamics (MD) simulations have been proposed.[21–23] One of the first methods based on classical molecular simulations was published by Seco and coworkers.[21] In this grid-based approach, an explicit restrained MD simulation of a protein is performed in the presence of a given concentration of isopropyl alcohol. The binding propensities of the probe at the protein surface are then back-computed to perform binding free energy calculations. A similar protocol was recently applied on different systems using several kinds of probes without any restraints on the protein.[23] The authors showed that probe molecules could induce both local and global structural rearrangements of the protein, leading to increases in target druggability. Nevertheless a frequent concern with these techniques is that the observed conformational changes reflect denaturation of the protein due to high probe concentrations. Thus judicious use of positional restraints is required to limit the occurrence of undesirable conformational changes. Also, probe diffusion necessary to compute binding propensities in buried cavities can be very slow with standard MD approaches. To overcome the limitations of current MD based druggability prediction methods, this report introduces the JEDI algorithm ('Just Exploring Druggability at protein Interfaces'). JEDI has been designed to evaluate protein druggability "on-the-fly" during MD simulations without any organic probes or protein restraints. The druggability function relies on a set of geometric parameters describing the volume, the enclosure and the hydrophobicity of a binding site. The JEDI scoring function is fast, continuous and differentiable. Accordingly, it can be used as a collective variable to bias MD simulations and enhance sampling of protein conformations. JEDI has been implemented in the software PLUMED 1.3 to enable metadynamics simulations and free-energy calculations with the most popular MD engines.[24] The methodology was

parameterized using the freely accessible Druggable Cavity Directory (DCD) dataset.[17] The sensitivity of the method to binding site conformational changes was tested with a compiled dataset of cryptic binding sites.

4.2 Materials & Methods

4.2.1 Overview of the JEDI approach

JEDI is a grid-based approach. The methodology includes three major steps (Figure 4.1A). First, a region of interest where the druggability evaluation will be conducted must be defined. This area can be located anywhere in the protein structure in principle, but in this report, efforts are focused on evaluating the druggability of known binding sites. Thus spatial regions to analyze were defined from the position of known ligands. A large 3D cubic grid with 1.5 Å spacing between grid points is initially positioned around the region of interest. Next, only grid points within 6 Å of one ligand atom were retained. All protein heavy atoms within 3 Å of a grid point are then selected for druggability calculations and this set of atoms is referred as the 'binding site region'. This setup is then followed by either a single point calculation or MD simulations with druggability evaluated at regular intervals in unbiased simulations, or at each time-step for MD simulations biased with the JEDI potential. Every druggability assessment requires that the 'activity' of all grid points is evaluated, with grid points classified as inactive, partially active or fully active according to their geometric position in the binding

site. Then, volume and hydrophobicity descriptors that depend on grid point activities and local geometric arrangements of protein atoms are computed in order produce a protein conformation dependent druggability score.

To avoid errors in the druggability predictions due to diffusion of the protein over the course of an MD simulation, the Cartesian coordinates of the grid points are re-evaluated prior to each druggability assessment. Firstly, the distance vector between the center of mass of the protein atoms in the binding site region in the conformation at the n-th step of the MD simulation ($\mathbf{r}_{com,t=n}$) and the initial protein conformation ($\mathbf{r}_{com,t=0}$) is evaluated. Then, the rotation matrix that best fits the protein backbone atoms of the binding site region onto their coordinates at t = 0 is computed using the Wolfgang Kabsch algorithm.[25] Finally, the resulting translation vector and rotation matrix are used to transform the grid point Cartesian coordinates at $t = 0$ into grid point Cartesian coordinates at $t = n$.

4.2.2 Datasets

Protein structures were taken from the Non Redundant Druggability Dataset (NRDD) in the DCD compiled by Schmidtke et al.[17] A set of 63 unique proteins has been used to parameterize the JEDI scoring function (Table 4.1). Each protein has been assigned by the authors of the original study an experimental druggability value from 1 to 10 (from less druggable to more druggable) according to its capability to bind a drug-like compound. The dataset can be further divided into three categories: non-druggable (DCDscore 1 to 4), difficult (DCDscore 5 to 7) and druggable (DCDscore 8 to 10). In order to benchmark JEDI against

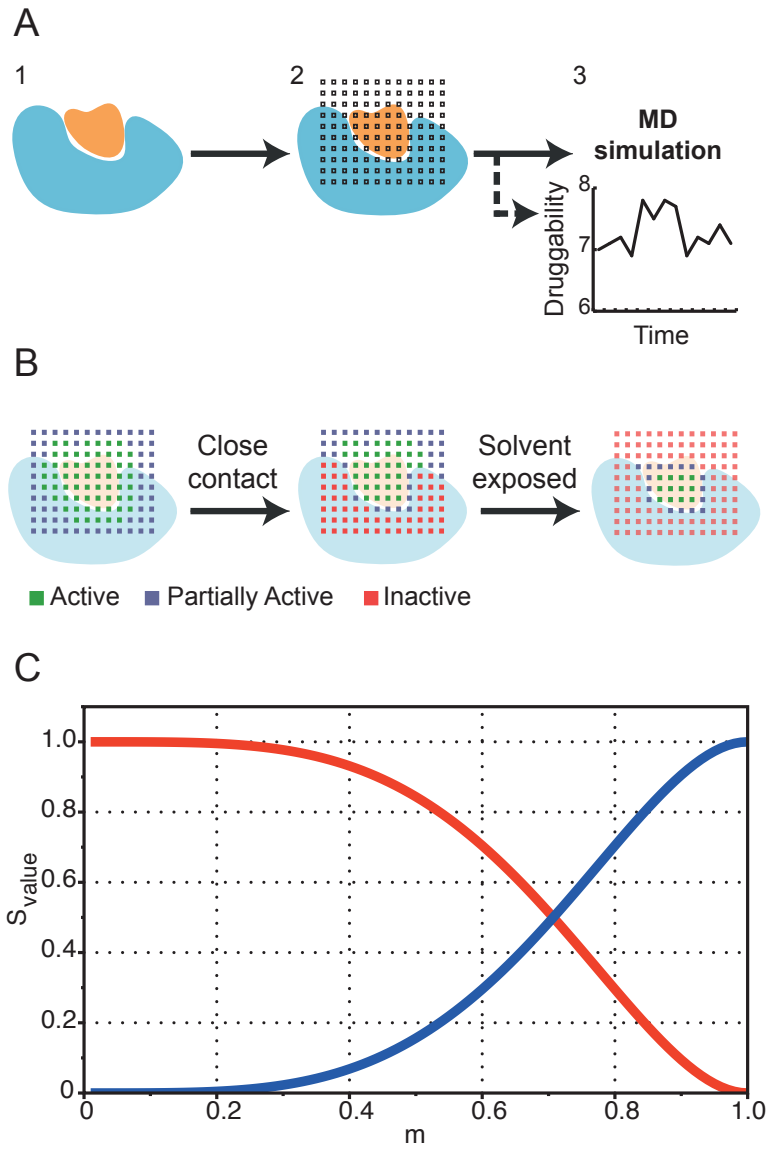

Figure 4.1: Overview of the JEDI protocol. A) The region of space for druggability assessment is determined and all atom models of the protein (and ligand if present) are prepared as for a conventional MD simulation (1). A grid with a 1.5 Å spacing is placed around the region of interest (2). A druggability assessment is performed either for the input structure only, or repeatedly over the course of an MD simulation(3). B) For every druggability evaluation, all grid points are assigned an initial activity according to their distance to the ligand in the input structure. Next, grid points overlapping with protein atoms in the binding site region are inactivated fully or partially. Finally, solvent exposed grid points are inactivated fully or partially. C) Graphical representation of the switching functions S_v^{on} (blue) and S_v^{off} (red) for $k = 1.0$ and $\Delta = 1.0$.

an existing methodology, druggability calculations were performed on the energy-minimized structures of the training dataset using the program fpocket.[16, 17] A detailed list of the dataset is given in Table 4.1, including druggability scores obtained with both approaches. A validation dataset, called the hidden pocket dataset, has also been compiled. Each protein in this dataset has two different structures that exhibit conformational variability in the binding site region that correlates with variations in the binding affinities of known ligands.

PDB code	DCD_{score}	$fpocket_{score}$	$JEDI_{score}$
1BMD	1	0.144	5.06
1BMQ	1	0.016	3.53
1CEN	1	0.087	4.75
1GYM	1	0.069	1.08
1I9Z	1	0.027	0.74
2F7F	1	0.097	2.29
2F94	1	0.585	2.94
2VZS	1	0.064	3.23
1NLJ	2	0.014	5.50
1DUD	3	0.602	4.96
1NNY	3	0.032	4.02
1G1F	3	0.250	4.26
1JF7	3	0.482	4.02
1ONZ	3	0.058	3.00
1AAX	3	0.600	3.93
1Q1M	3	0.300	4.60
2GSS	3	0.731	4.40
2COI	3	0.113	4.71

1VQ1	4	0.684	6.73
6PAH	4	0.802	5.97
1AJ6	6	0.238	5.50
1FCM	6	0.008	1.94
1HW9	6	0.185	3.92
1JRP	6	0.498	6.53
1KIJ	6	0.119	6.65
1PBF	6	0.033	5.26
1QZR	6	0.257	5.77
1R55	6	0.083	2.51
1T48	6	0.489	7.46
2PK4	6	0.419	6.39
2RMC	6	0.564	6.66
1EVE	7	0.751	6.38
1EZQ	7	0.206	5.26
1KI2	7	0.409	6.75
1E9X	7	0.819	8.29
1TBF	7	0.671	7.32
1W22	7	0.516	3.06
1C14	8	0.831	8.07
1TH6	8	0.706	4.22
1DYR	8	0.758	6.78
1E1X	8	0.729	7.18
I1EP	8	0.804	9.20
1PXX	8	0.885	8.33
1TBB	8	0.307	6.55
1U72	8	0.757	6.33

2BL9	8	0.809	5.61
3D4S	8	0.693	7.07
1DTL	9	0.822	10.4
1ZGY	9	0.809	9.93
1FK9	9	0.843	9.71
1HSH	9	0.603	8.74
1IHI	9	0.899	7.30
1KV1	9	0.753	7.78
1PWL	9	0.852	8.63
2H79	9	0.757	8.83
3D1X	9	0.735	8.11
1A28	10	0.955	8.77
1ERR	10	0.901	9.34
1LHU	10	0.811	7.57
1NHZ	10	0.963	8.68
2OAX	10	0.943	10.09
2Q7J	10	0.940	8.94
3CAJ	10	0.542	6.64

Table 4.1: Details of the dataset used for the parameterization of JEDI.

4.2.3 JEDI scoring function

The JEDI druggability score is calculated as a linear combination of two partial-least squared derived descriptors reflecting the volume, and the hydropho-

bicity (Equation 4.1).

$$JEDI_{score} = V_{druglike} \left(\alpha V_a + \beta H_a + \gamma\right) \tag{4.1}$$

where $V_{druglike}$, V_a and H_a represent respectively the drug-like volume descriptor, the pocket volume descriptor and the pocket hydrophobicity. α, β and γ are constants of the model derived by multiple linear regressions against a training set. All the descriptors presented below are based on cubic spline functions such that the $JEDI$ potential is continuous and twice differentiable. Two forms of cubic spline functions have been used operating on variables v and k (Figure 4.1 C). The first one turns 'off' with v starting at k at v_{min}, reaching 0 at $v_{min} + \Delta$ (Equation 4.2).

$$S_v^{off}\left(k, v_{min}, \Delta\right) = \begin{cases} k \text{ if } m < 0 \\ k\left[\left(1 - m^2\right)^2 \left(1 + 2m^2\right)\right] \text{ if } 0 \leq m \leq 1 \\ 0 \text{ if } m > 1 \end{cases} \tag{4.2}$$

where $m = \frac{v - v_{min}}{\Delta}$. The second form turns 'on' the variable S from 0 to k along an interval Δ (Equtation 4.3).

$$S_v^{on}\left(k, v_{min}, \Delta\right) = \begin{cases} 0 \text{ if } m < 0 \\ k\left[1 - \left(1 - m^2\right)^2 \left(1 + 2m^2\right)\right] \text{ if } 0 \leq m \leq 1 \\ k \text{ if } m > 1 \end{cases} \tag{4.3}$$

The active volume descriptor V of the binding site is given by Equation 4.4:

$$V = \sum_{i=1}^{N} a_i V_g \tag{4.4}$$

where N is total number of grid points, V_g is the volume of space covered by a grid point. To capture the shape of the pocket, each grid point is assigned an activity score a_i between 0 and 1 (inactive to active), according to its geometric position inside the binding pocket (Equation 4.5).

$$a_i = S^{off}_{BS_i}\left(1.0, BS_i, \Delta BS\right) S^{on}_{mind_i}\left(1.0, CC_{mind}, \Delta CC\right) S^{on}_{exposure_i}\left(1.0, E_{min}, \Delta E\right) \tag{4.5}$$

The first term of Equation 4.5 gradually turns off grid points according their distances from the region of interest. This term is optional, but is useful to ensure that fluctuations in druggability scores are not unduly influenced by conformational changes that are remote from the protein region of interest. The minimum distance BS_i between a grid point i and the M atomic coordinates defining the binding site region is calculated as

$$BS_i = \frac{\theta}{\ln\left(\sum_{j=1}^{M} \exp\left(\frac{\theta}{\|\mathbf{r}_{ij}\|}\right)\right)} \tag{4.6}$$

With $\theta = 50.0$ Åand $\mathbf{r}_{ij} = \mathbf{r}_{gi} - \mathbf{r}_{pj}$, where $\mathbf{r}_{gi}$ and $\mathbf{r}_{pj}$ are respectively the position vectors of grid point i and protein atom j belonging to the binding site region. The second term in Equation 4.5 causes grid points that overlap with protein atoms to be gradually inactivated (Figure 4.1B). The minimum distance $mind_i$ between grid points and protein atoms is calculated with an equation similar to Equation 5.16. The third term in Equation 4.5 gradually inactivates solvent

exposed grid points (Figure 4.1B).

$$exposure_i = \sum_{k=1}^{N} [S^{off}_{mind_k}\left(1.0, CC2_{min}, \Delta CC2\right) S^{on}_{\|\mathbf{r}_{ik}\|}\left(1.0, GP1_{min}, \Delta GP1\right) S^{off}_{\|\mathbf{r}_{ik}\|}\left(1.0, GP2_{min}, \Delta GP2\right)] \tag{4.7}$$

where $CC2_{min}/\Delta CC2$ control the distance below which a grid point is considered as interacting with the protein. $GP1_{min}/\Delta GP1$ and $GP2_{min}/\Delta GP2$ are used to select grid points at a given distance interval from the grid point i in order to penalize solvent exposed grid points. With the default values presented in Table 4.2, a maximum of 44 grid points can be selected around a given grid point i and the maximum value of $exposure_i$ is 23.97 with the present parameterization.

Symbol	Definition	Value
$V_{druglike}$	drug-like volume descriptor	0 to 1
V_a	pocket volume descriptor	$[0, \infty]$
H_a	pocket hydrophobicity descriptor	0 to 1
V	active volume	$[0, \infty]$
a_i	activity of the grid point i	0 to 1
H_i	hydrophobicity of the grid point i	0 to 1
M_{apolar}	number of C and S atoms to calculate $JEDI_{score}$	$[0, \infty]$
$apolar$	number of C and S atoms surrounding the grid point i	$[0, \infty]$
M_{polar}	number of O and N atoms to calculate $JEDI_{score}$	$[0, \infty]$
$polar$	number of O and N atoms surrounding the grid point i	$[0, \infty]$

Table 4.2: List of variables used to compute $JEDI_{score}$.

The active volume V is then converted in a pocket volume descriptor V_a using Equation 4.8.

$$V_a = \frac{V}{V_{max}} \tag{4.8}$$

where V_{max} is the maximum active volume descriptor. This constant was set to be equal to the maximum active volume V calculated for protein binding sites in the 'druggable' category of the DCD dataset. Accordingly, a cavity presenting

the characteristics of a typical small-molecule binding site will have a typical V_a value in the interval $[0.0, 1.0]$. In order to penalize overly large or overly small cavities that are not suitable for drug-like small molecules, the descriptor $V_{druglike}$ is also computed with Equation 4.9.

$$V_{druglike} = S_V^{off}\left(1.0, V_{max}, \Delta V_{max}\right) S_V^{on}\left(1.0, V_{min}, \Delta V_{min}\right) \tag{4.9}$$

where V_{min} is equal to 0 Å^3 by default. Analysis of pockets from the DCD dataset suggested a ΔV_{min} value of 36 Å^3. For simplicity, the same value was used for Δ_{max}. The effect is that cavities that differ substantially in active volume from those present in the training set will have a low value of $V_{druglike}$. In turn this will assign a low $JEDI_{score}$ to cavities that differ markedly from the training set.

The active grid hydrophobicity function captures the average hydrophobicity of the active grid points and is given by equtation 4.10:

$$H_a = \frac{1}{V}\sum_{i=1}^{N}\left(H_i a_i\right) \tag{4.10}$$

where the hydrophobicity score H_i of the grid point i is calculated as

$$H_i = \frac{apolar_i}{apolar_i + polar_i} \tag{4.11}$$

where $apolar_i$ and $polar_i$ are respectively the number of apolar (carbon and sulfur) and polar (oxygen and nitrogen) protein atoms within the distance r_{hydro} defined by equations 4.12 and 4.13:

$$apolar_i = \sum_{j=1}^{M_{apolar}} S_{\|\mathbf{r}_{ij}\|}^{off}\left(a_i, r_{hydro}, \Delta r_{hydro}\right) \tag{4.12}$$

$$polar_i = \sum_{j=1}^{M_{polar}} S^{off}_{\|\mathbf{r}_{ij}\|}(a_i, r_{hydro}, \Delta r_{hydro}) \tag{4.13}$$

4.2.4 JEDI optimization

The parameters of the JEDI model were optimized using the python module PyEvolve.[26] After investigation, only the CC_{mind}, ΔE, $\Delta CC2$ and r_{hydro} variables presented in the Table 4.4 were selected for optimization using a range of physically plausible values (Table 4.3).

Symbol	Values	Units
CC_{mind}	1.6, 1.8, 2.0, 2.2, 2.4, 2.6	Å
ΔE	1, 3, 5, 7, 9, 11, 13	-
$\Delta CC2$	0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6	Å
r_{hydro}	3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0	Å

Table 4.3: Range of values used for JEDI optimization.

An elitist genetic algorithm was then iterated for 50 generations on a population of 40 individuals. All individuals consisted of a combination of four parameters. The value of each parameter was randomly selected according to the range of values presented in Table 4.3. The fitness function was defined to maximize the r^2 of $JEDI_{score}$ vs DCD_{score} values after a Partial Least Squares regression. The convergence of the r^2 was manually verified and the corresponding individual was selected. Uncertainties in the $JEDI_{score}$ parameters were determined with 100 iterations of bootstrapping using a split of 0.7/0.3 for the training and validation sets.

Symbol	Definition	Value
α	PLS derived volume coefficient	5.31
β	PLS derived hydrophobicity coefficient	24.29
γ	PLS derived constant according to αandβ	-13.39
V_g	grid resolution	1.5 Å^3
CC_{mind}	distance below which a grid point is fully in close contact with the protein	2.0 Å
ΔCC	distance interval over which a grid point is in partial contact with the protein	0.5 Å
E_{min}	minimum number of grid points between a distance of 2.5 Å and 3.5 Å from a grid point i interacting with the protein	10
ΔE	interval over which a grid point is considered as buried in the cavity	3
BS_{min}	minimum distance between a grid point and binding site atoms below which the maximal activity is fixed to 1	2.0 Å
ΔBS	distance interval over which the maximal activity is fixed to 0	6.0 Å
θ	constant used for minimum distance calculation	50.0 Å
$CC2_{min}$	minimum distance below which a grid point is overlapping the protein (for enclosure calculation)	0.15 Å
$\Delta CC2$	distance interval over which a grid point is in partial contact with the protein (for enclosure calculation)	0.14 Å
$GP1_{min}$	distance above which a grid point is considered for enclosure calculation	2.5 Å
$\Delta GP1$	distance interval over which a grid point is in partial contact with the protein	0.5 Å
$GP2_{min}$	distance below which a grid point is fully in close contact with the protein	3.0 Å
$\Delta GP2$	distance interval over which a grid point is in partial contact with the protein	0.5 Å
r_{hydro}	distance below which a grid point is fully in close contact with the protein (for hydrophobicity calculation)	4.0 Å
Δr_{hydro}	distance interval over which a grid point is in partial contact with the protein (for hydrophobicity calculation)	0.5 Å
V_{max}	volume below which $V_{druglike}$ is equal to 1	316 Å^3
ΔV_{max}	volume interval over which $V_{druglike}$ goes from 1 to 0	36 Å^3
V_{min}	volume below which $V_{druglike}$ is equal to 0	0.0 Å^3
ΔV_{min}	volume interval over which $V_{druglike}$ goes from 0 to 1	36 Å^3

Table 4.4: List of constants used to compute $JEDI_{score}$.

4.3 Results

4.3.1 Choice of descriptors

The druggability score of the JEDI methodology is based on a linear combination of structural descriptors characterizing the volume and the hydrophobicity of a cavity. The choice of those collective variables were influenced by the literature.[6, 12, 13, 17, 18, 27] A rule-based method published by Perola et al. suggested five suitable descriptors: volume, depth, enclosure, percentage of charged residues and hydrophobicity. These descriptors summarize a general consensus fairly well.[28] After investigation, only two descriptors, the active volume and the hydrophobicity, have been retained. An early version of JEDI also included a descriptor capturing the degree of 'buriedness' of the binding site. The buriedness, as described by Volkamer et al., was captured as the ratio between the number of hull grid points in contact with the protein surface and the total number of hull grid points.[27] After preliminary investigations, this descriptor was not found to contribute significantly to the druggability prediction. This is likely because the current definition of the active volume descriptor is penalizing solvent-exposed grid points and thus already accounts for buriedness. Consequently, shallow solvent exposed cavities have a lower active volume descriptor than buried enclosed closed cavities. The results depicted in Figure 4.2 demonstrate that higher $JEDI_{score}$ values do correlate with a larger binding site active volume V and a larger hydrophobicity descriptor H_a.

Since the publication of the first large scale classification of protein binding

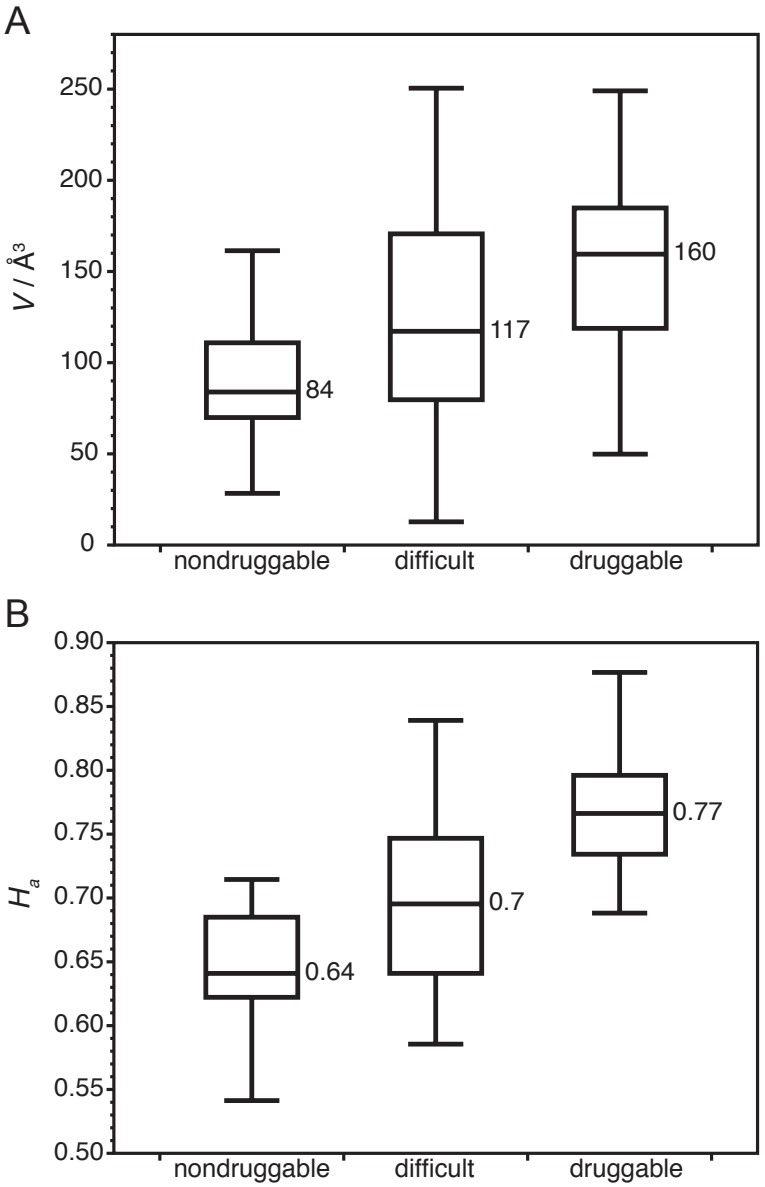

Figure 4.2: Boxplots of values of the (A) active volume V and (B) hydrophobicity descriptor H_a for the nondruggable, difficult and druggable systems of the training set. The box is defined using the first and the third quartile while the bar indicates the median. The edges of the boxplot represent the minimum and the maximum value observed for each category.

sites by An et al,[29] numerous studies have been conducted in the field of pocket detection and analysis to improve understanding of the physicochemical properties that underlie protein-ligand interactions.[6, 17, 18, 27] The average volume of a druggable binding site was evaluated around 600 $Å^3$,[29] with maximum values around 900-1200 $Å^3$.[27, 28] These estimates are in line with those computed with JEDI; the average volume of a binding site represented by the total number of active and partially active ($a_i > 0$) grid points was found to be 496 ± 202 $Å^3$ with a maximum value of 1019 $Å^3$. The results shown in Figure 4.2A depict the distribution of active volume (V) values for different categories of protein binding sites. As the active volume is the sum of the grid point volumes weighted by their activity, it is in general much smaller than the volume of the binding site. An average value for the whole dataset is $V = 125 \pm 60$ $Å^3$.

The JEDI hydrophobicity descriptor shares similarities with the descriptor used by by Eyrisch et al.[30, 31] In accordance with previous literature studies, druggable binding sites tend to have higher average hydrophobicity values ($H_a = 0.72 \pm 0.03\sigma$) than non-druggable binding sites ($H_a = 0.60 \pm 0.04\sigma$). This descriptor was found to be the most significant contribution to the $JEDI_{score}$ values with a weight β almost five times larger than the α volume coefficient (Table 4.2). This observation is in good agreement with the literature, where the apolar character of a cavity is usually the most important structural descriptor for druggability assessment.[13, 27]

4.3.2 Druggability scoring of diverse protein structures

The JEDI parameters were first optimized using multiple linear regressions and the elitist selection variant of the genetic algorithm methodology implemented in the python module PyEvolve.[26] JEDI druggability scores obtained at the end of the process are shown in Figure 4.3A. For comparison, fpocket was used to calculate the druggability score of each protein in the training dataset (Figure 4.3B). The results suggest that JEDI predictions are slightly more accurate than those obtained using fpocket with a r^2 of 0.63 ± 0.11 and 0.52 ± 0.13 respectively. Closer inspection of Figure 4.3A shows that JEDI discriminates fairly well undruggable sites from druggable sites, but proteins in the difficult category show a large scatter in $JEDI_{score}$ values. Clearly, the precise 'experimental' DCD druggability score to assign to a protein can be debated, and this must be kept in mind when calibrating computational methods against this dataset. Additional tests were conducted by positioning the grid on buried or solvent exposed regions of the protein Malate Dehydrogenase (1BMD), where no apparent pockets were observed. The resulting $JEDI_{score}$ values where invariably lower than 1.5.

Detailed structural analyses of accurate and inaccurate druggability predictions for representatives druggable and non-druggable protein binding sites is useful to characterize the strengths and weaknesses of the present approach. Four representative structures were chosen for this purpose (Figure 4.4), and JEDI descriptor values for these structures are shown in Table 4.5.

Figure 4.4A represents the binding site of a malate dehydrogenase in complex with the coenzyme NAD (PDB 1BMD). This enzyme has been classified as nondruggable due to the difficulty of finding a drug-like compound able to compete

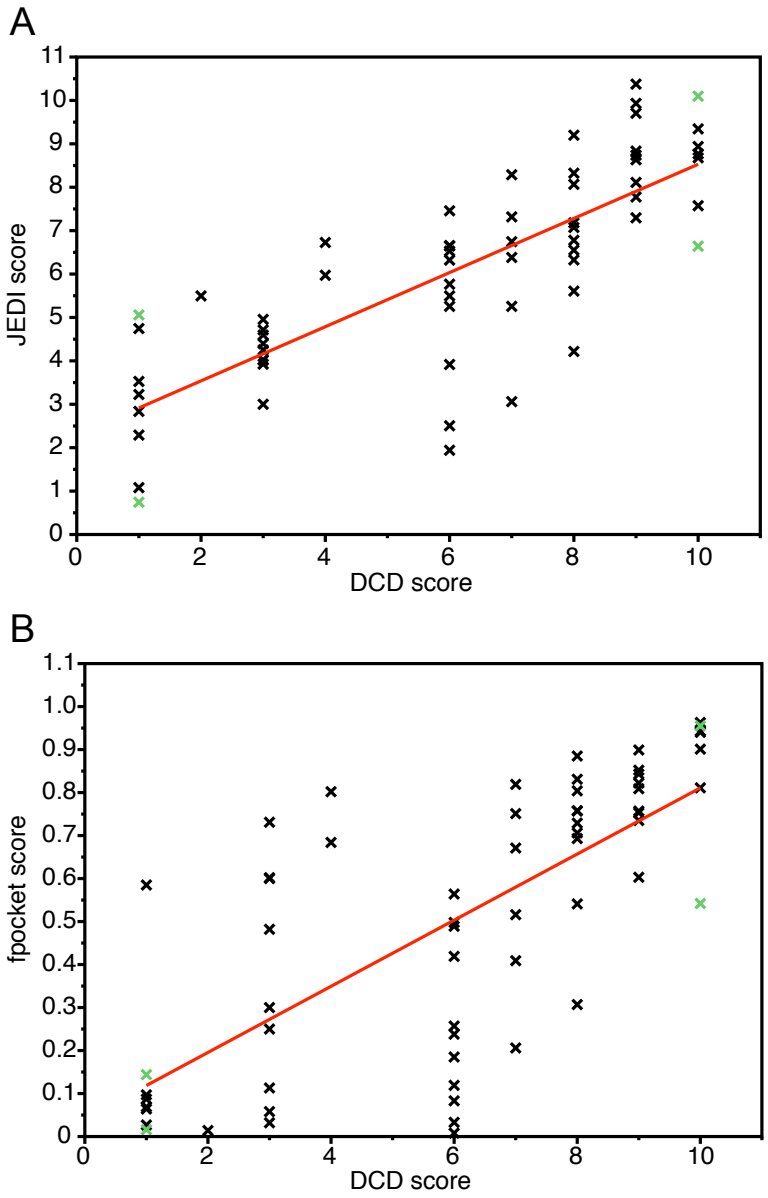

Figure 4.3: The correlation of computed druggability scores with DCD database druggability scores. A) Results for JEDI scores of the DCD training set. B) Results for fpocket scores of the DCD training set. Proteins discussed in the text, Figure 4.4 and Table 4.5 are represented with green crosses.

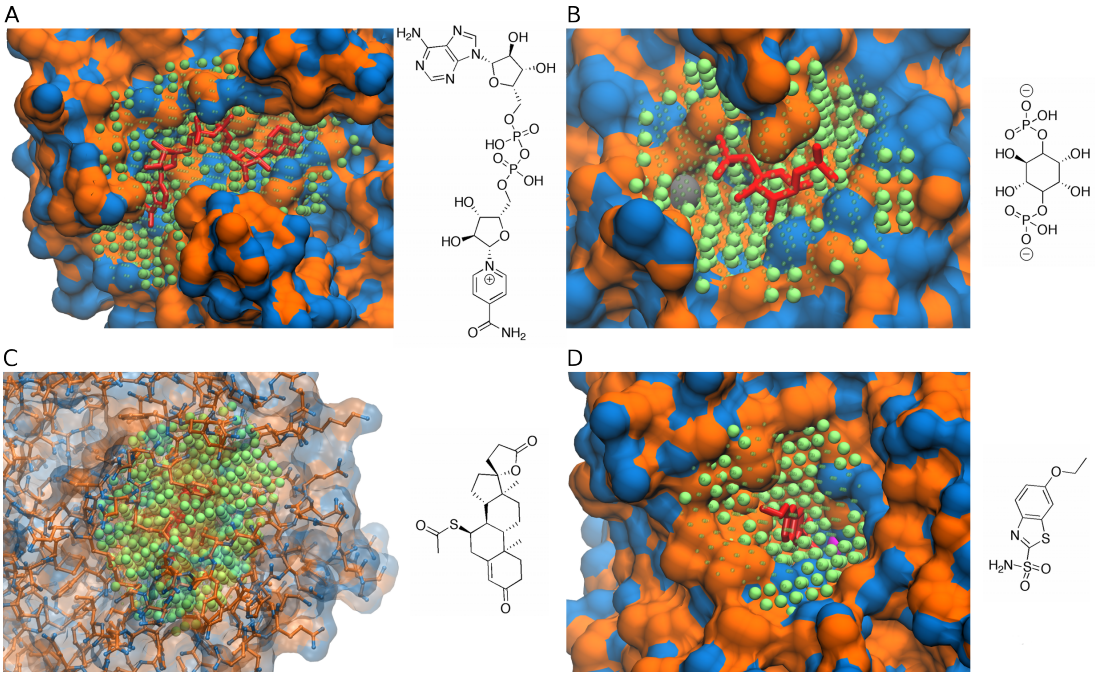

Figure 4.4: The relationship between JEDI druggability scores, binding site descriptors and ligand structures. A) Malate dehydrogenase is a nondruggable target predicted to have an intermediate druggability score. It is in a complex here with the coenzyme NAD (PDB 1BMD). B) IP phosphatase is a nondruggable binding site that is predicted to have a low druggability score. It is here in a complex with inositol(1,4)-bisphosphate and a calcium ion (PDB 1I9Z). C) Mineralocorticoid receptor is a druggable target that is predicted to have a high druggability score. It is here in a complex with spironolactone (PDB 2OAX). D) Carbonic anhydrase II is a druggable target that is predicted to have a low druggability score. It is here in a complex with ethoxzolamide and a zinc ion (PDB 3CAJ). The protein surface has been colored according to polar (blue) and apolar (orange) atoms. The 3D ligand conformations are represented in red licorice. Green dots symbolize grid points, and grid points with activity values $a_i > 0$ are depicted with smaller spheres. Calcium and zinc ions are respectively represented as grey and pink van der Waals spheres. Pictures were prepared using the software VMD.

with NAD for access to the binding site. The binding affinity of several known nucleotide inhibitors have been previously determined by enzymatic assays.[32] The best competitive inhibitor is the cyclic nucleotide cAMP, presenting a K_i value 560 nM. If this system is clearly evaluated as nondruggable by fpocket (score = 0.11), it remains challenging for other methodologies such as the NMR-based approach developed by Hadjuk and coworkers, which predicts the cavity as having an intermediate druggability. This is in line with the observed $JEDI_{score}$ value for this system (5.1). The relatively high $JEDI_{score}$ is largely due to the relative large active volume V of the binding site (157 $Å^3$), which is in the range of V values typical for druggable sites (Table 4.5, first row). Thus, that malate dehydrogenase is not considered druggable in practice may be more a reflection of the difficulty for a drug-like molecule to compete with NAD at a ca. 300 μM expected intracellular concentration in mammalian cells,[33] rather than the occurrence of an unusually polar or shallow binding site.

An example of a correct nondruggable prediction is depicted in Figure 4.4B for the binding site of Inositol Polyphosphate (IP) phosphatase.[34] In addition to a small active volume due to a poor degree of enclosure, this small pocket presents a very low hydrophobicity score (Table 4.5, second row). This is mainly because of a Calcium ion in the binding site.

A correctly predicted druggable cavity is shown in figure 4.4C. This mostly apolar well-enclosed pocket corresponds to the binding site of the S810L mutant mineralocorticoid receptor interacting with spironolactone (Table 4.5, third row).[35] This inhibitor has shown IC50 values in the range of $1.6 - 60$ nM in a cell-based luciferase reporter assay.[36]

Lastly, Figure 4.4D depicts a druggable binding site that is incorrectly predicted to be 'difficult' to target. In addition to a high polarity caused by the

presence of a zinc ion buried in the pocket, the binding site of carbonic anhydrase II is particularly small.[37] Most successful carbonic anhydrase inhibitors exploit direct interactions with the buried Zinc ion. The present version of JEDI does not account for potentially favorable metal-ligand interactions and this explains the discrepancy between the $JEDI_{score}$ and DCD_{score} values (Table 4.5, fourth row).

Protein	V / Å^3	H_a	$JEDI_{score}$	DCD_{score}
Malate dehydrogenase	157	0.64	5.1	1
IP phosphatase	34	0.57	0.7	1
Mineralocorticoid receptor	236	0.80	9.7	10
Carbonic anhydrase II	85	0.76	6.6	10

Table 4.5: JEDI descriptor values for the structures depicted in figure 4.4

4.3.3 Sensitivity to minor structural variations, and performance

A potential concern at the outset of the project was that $JEDI_{score}$ values would be unduly sensitive to minor structural variations that are typically observed when crystal structures of the same protein are solved and refined independently. A major motivation for the development of JEDI was to observe variability in $JEDI_{score}$ between different structures of the same protein, only when conformational changes relevant for drug design are observed (e.g. a side-chain flip). This feature requires a subtle balance, on the one hand the methodology should not be too sensitive to very minor structural changes, but on the other hand it should be sufficiently sensitive to capture a fluctuation in druggability if the rearrangement is significant. The strategy here adopted was to evaluate the sensitivity of the $JEDI_{score}$ values for comparable conformations of the same protein interacting

with different ligands. The structural similarity was quantified by means of RMSD calculations on the backbone and $textC_\beta$ atoms of the binding site atoms of each protein. Selected proteins for which RMSD values of the different structures were less than 0.5 Å were retained for further analysis. Additionally, visualization of the binding sites confirmed that there was no noticeable difference in binding site conformation between the different selected structures. Figure 4.5A shows the distribution of $JEDI_{score}$ values obtained by this analysis for a representative protein taken from the 'nondruggable', 'difficult' and 'druggable' categories of the DCD dataset. Although small fluctuations in $JEDI_{score}$ are observed in the case of the difficult and the druggable binding site, the results suggest nevertheless a good reproducibility and robustness to insignificant structural changes. By contrast the fpocket methodology sometimes exhibits substantial variations in druggability that complicates interpretation of the scores (Figure 4.5B). As an additional test of sensitivity, the dependence of the $JEDI_{score}$ values on the initial placement of the grid was assessed by evaluating the druggability of the same protein after translations of grid point coordinates by up to ± 0.5 Å in the x, y, and z directions in Cartesian space. The druggability predictions were found to be quite insensitive to such translations, with fluctuations in the $JEDI_{score}$ values in the range of 0.1.

Next the computational cost of the JEDI calculations was assessed. An important consideration is that the calculations should not slow down too much molecular dynamics simulations. Benchmarks are shown in Table 4.6. If JEDI is used to monitor druggability values on the fly during an MD simulation, then it is not necessary to evaluate druggability at every time-step, as snapshots between successive times-steps are highly correlated. With druggability evaluation every 1 ps the time incurred is negligible, unless the MD simulation is parallelized across multiple processors. Likewise, single-point druggability estimates of a

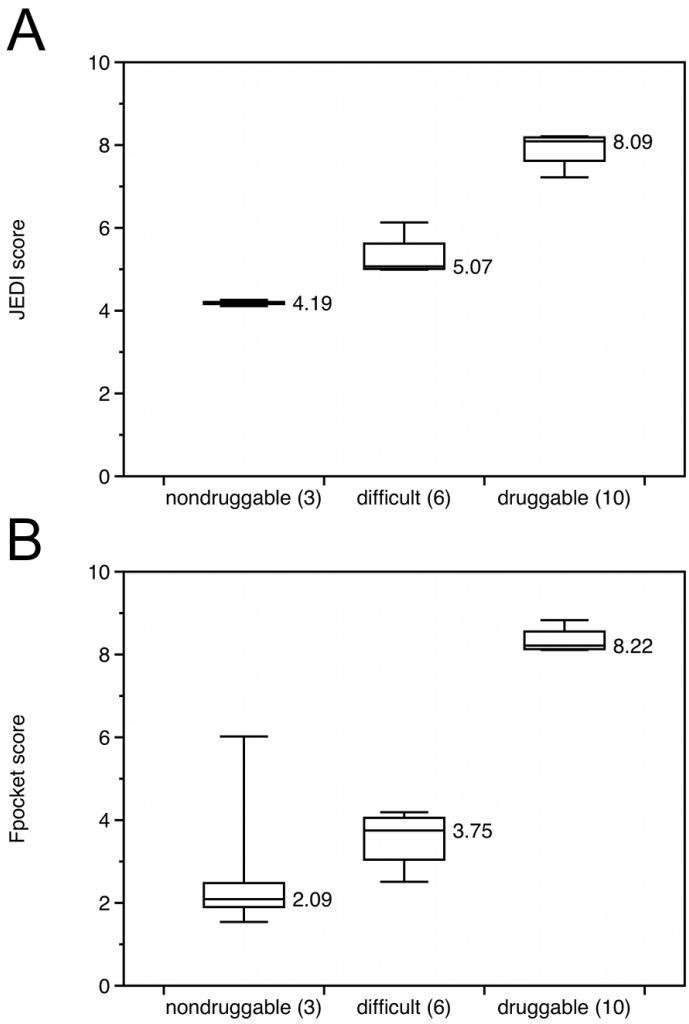

Figure 4.5: The sensitivity of druggability scores to small structural differences. The boxplots illustrate the fluctuations of the, (A) JEDI and, (B) fpocket druggability scores obtained from several highly similar conformations of a binding site for three different proteins. The DCD druggability score of each protein is given in parenthesis in the x-axis. The nondruggable, difficult and druggable systems selected for druggability assessment were respectively the dUTPase (PDB codes 1DUD, 1RN8, 1RNJ, 1SEH, 1SYL, 2HR6, 2HRM), the Kringle 1 domain of human plasminogen (PDB codes 1CEA, 1CEB, 2PK4, 1HPK) and the human sex hormone-binding globulin (PDB codes 1LHN, 1LHU, 1LHV, 1LHW). For the sake of consistency, only protein structures presenting a binding site identified by fpocket were selected.

protein structure are far faster than alternative methodologies that take seconds to minutes.[15, 17, 27] The implementation of MD simulation protocols biased with JEDI requires a druggability calculation at each time-step. In this case the performance loss is approximately a factor of 1.4 to 2.7, depending on the number of processors used to speed-up the evaluation of the non-bonded energies. Evidently, further gains in efficiency could be gained by parallelizing key subroutines in the JEDI code. The relative efficiency is also influenced by the choice of an implicit solvent model for this study, which dramatically speeds up the evaluation of non-bonded energies. Overall, the performance was deemed acceptable, given scope for future improvements.

System	Number processors	MD	MD/JEDI (monitor mode)	MD/JEDI (bias mode)
VHL	1	1.3	1.3	0.9
	2	2.5	2.5	1.1
	4	3.1	3.0	1.1
hPNMT	1	0.5	0.5	0.4
	2	0.8	0.8	0.5
	4	1.6	1.3	0.6

Table 4.6: JEDI performance in ns/day for VHL (2278 atoms) and hPNMT (4057 atoms). The results were obtained using a cut-off of 20 Å for the neighbor list, and 100 ps simulations on an Intel Xeon E3-1270 v3 (3.5GHz) processor.

4.3.4 Application to a hidden pockets dataset

Validation of the methodology was pursued by analysis of a set of six proteins known to adopt distinct binding site conformations in the presence of different ligands (Figure 4.6). In each instance, two conformations for each protein were selected for druggability assessments. Protein structures were aligned

and a grid defined from the largest ligand was used to compute a JEDIscore value for both conformations. In all instances the ligand atoms were ignored for druggability calculations. The results of this analysis are shown in Table 4.7. Human phenylethanolamine N-methyltransferase (hPNMT) is an enzyme involved in the synthesis of epinephrine from norepinephrine using the cofactor S-adenosyl-L-methionine to methylate the primary amine of noradrenaline. Two different hPNMT inhibitors, 1 and 2, have been reported to inhibit the enzyme with K_i values of 0.28 μM and 0.063 μM respectively (radiochemical assay).[38] It has been shown that these two ligands bind to different conformations of the hPNMT binding site (Figure 4.6A). Both compounds engage in significant hydrophobic interactions, but the larger ligand (2) positions a p-chlorophenyl group in a cavity that is hidden in the hPNMT/1 complex. Formation of the enlarged cavity in hPNMT/2 necessitates the rearrangement of the side-chain Lys57, as well as a small displacement of helix α3. The JEDI calculations were able to capture a favorable increase in druggability of ca. 0.8 units for the protein conformation seen in hPNMT/2 in comparison with hPNMT/1. The change in druggability is due to a favorable increase in both V and H_a (Table 4.7, first row).

The von Hippel-Lindau protein (pVHL) forms a complex with the proteins CUL2, Elongin B and C, and Rbx1. This complex is involved in the ubiquitination of the transcription factor hypoxia-inducible factor (HIF-1α), leading to proteasome-mediated degradation of HIF-1α.[39] Small molecules 3 and 4 have been reported to inhibit interactions between pVHL and HIF-1α with K_d values of 86.1 μM and 27.7 μM respectively (fluorescence polarization assay).[20] The ligands occupy the same binding site, but a different orientation of Arg107 is observed, giving rise to a slightly more enlarged cavity in VHL/4 (Figure 4.6B). This translates into a slightly higher $JEDI_{scorevalue}$ for VHL/4 over VHL/3. This is because repositioning of Arg107 increased the value of H_a in VHL/4. However

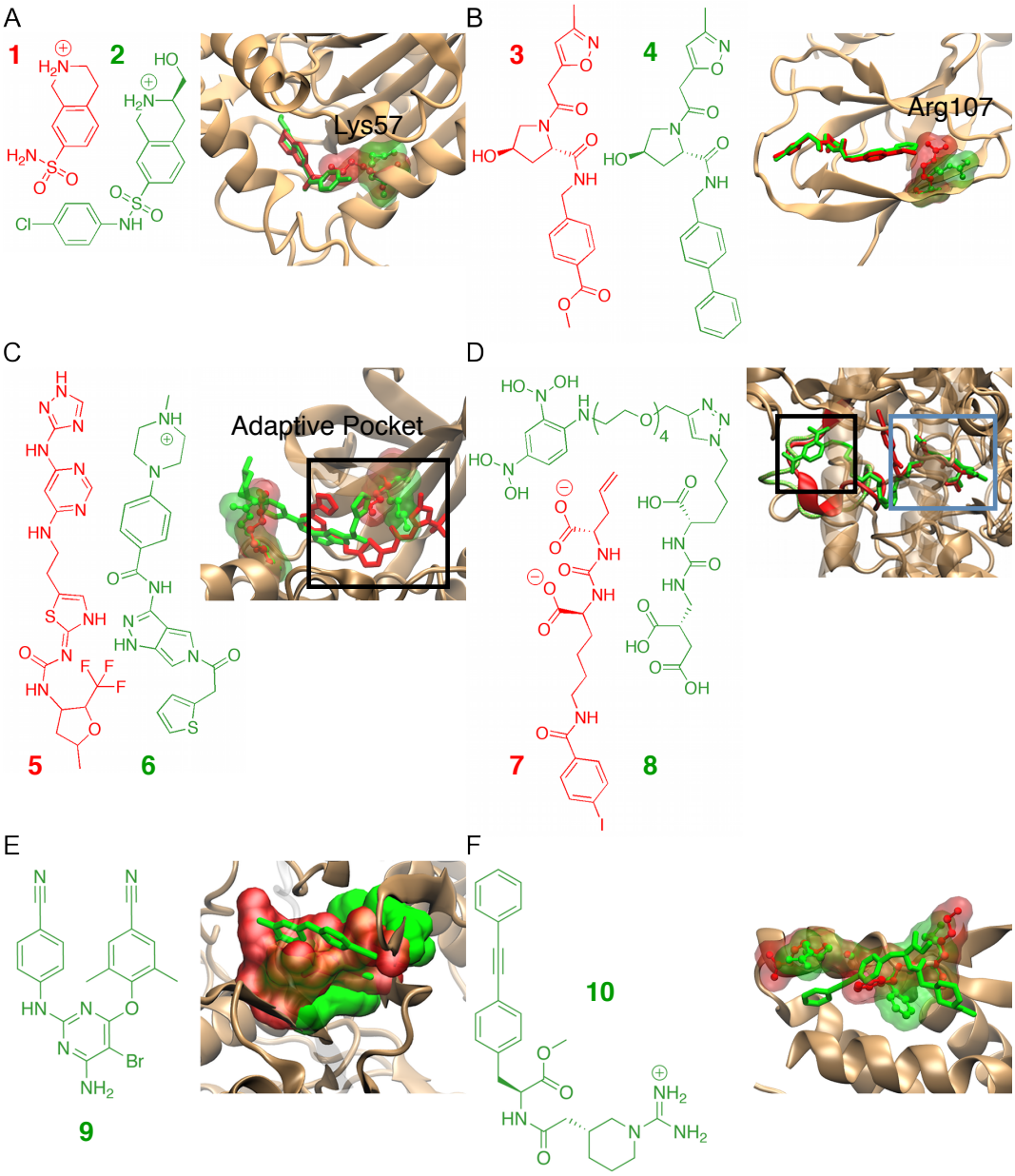

Figure 4.6: Conformational variability of the hidden pocket dataset. (A) hPNMT in complex with 1 or 2, (B) VHL in complex with 3 or 4, (C) PLK-1 in complex with 5 or 6, (D) PSMA in complex with 7 or 8, (E) HIV-1 in complex with 9, (F) IL-2 in complex with 10. Protein regions that are similar in both conformations are represented in brown. 3D structures of the ligands are displayed in licorice. Pictures were prepared using the software VMD.

this is partially offset by a decrease in V_a. This is because the displacement of Arg107 exposes more grid points to the solvent, and as a consequence, grid points previously fully active become partially active (Table 4.7, second row).

Ligand	Protein	PDB code	$JEDI_{score}$	V / Å^3	H_a	$V_{druglike}$
1	hPNMT	1HNN	8.4	259	0.72	1.0
2		2G8N	9.2	276	0.74	1.0
3	VHL	3ZTD	8.2	118	0.80	1.0
4		3ZTC	8.5	114	0.82	1.0
5	PLK-1	2OWB	8.9	247	0.74	1.0
6		3DB6	8.1	223	0.72	1.0
7	PSMA	3IWW	0.0	493	0.54	0.0
8		2XEG	4.7	341	0.55	0.8
-	HIV-RT	1DLO	8.5	192	0.76	1.0
9		3M8P	9.6	213	0.78	1.0
-	IL-2	1M47	7.3	77	0.80	1.0
10		1M48	6.2	78	0.75	1.0

Table 4.7: JEDI descriptor values for the hidden pocket dataset.

Serine/threonine-protein kinase or polo-like kinase 1 (PLK-1) is an enzyme involved in the regulation of cell division., The PLK-1 inhibitor 5 binds with an IC50 = 730 nM (fluorescence polarization assay) to the ATP binding site, and also to a subpocket that has been called the adaptive pocket, whereas the inhibitor 6 shows an IC50 of 530 nM (kinase enzymatic assay) and binds to the native purine-pocket of the active site (Figure 4.6C).[40, 41] However the larger active volume observed in the PLK-1/5 bound conformation is mainly due to active grid points around the methylpiperazine moiety of 5. These grid points are inactive in the PLK-1/6 complex because they are too solvent exposed. The adaptive pocket seen in PLK-1/6 is predicted to be less druggable than the native pocket seen in PLK-1/5 by ca. 0.8 units (Table 4.7, third row).

Prostate specific membrane antigen (PSMA) is a glycoprotein overexpressed as a homodimer in many forms of prostate cancer. Compound 7 is an example of a

first generation of PSMA inhibitors that bind the very polar binding site of PSMA with a K_i of 11 nM (fluorescence-based NAALADase assay).[42] More recently, compounds belonging to the class of antibody recruiting small molecules targeting prostate cancer (ARM-P) have been reported, and compound 8 binds PSMA with a K_i of 0.02 nM (enzymatic assay).[43, 44] A crystallographic structure of the PSMA/8 complex revealed that 8 binds to an open PSMA conformation that was not observed in the PSMA/7 complex. The large difference in binding affinities between 7 and 8 appears to be well reproduced by a large difference in $JEDI_{score}$ values (Table 4.7, fourth row). However in this instance the active volume V is much larger than for a typical small molecule binding site and as a consequence the druggability score is strongly penalized by $V_{druglike}$. This indicates that the predictions should be treated with care as the binding site differs substantially from those present in the training set. Compound 8 is unusual because it is made of a long flexible linker connecting a moiety positioned in the buried PSMA active site (Figure 4.6D blue square), and another moiety positioned in the arene binding site at the protein surface (Figure 4.6D black square). The JEDI analysis was therefore repeated by splitting the initial grid in two regions in order to predict the druggability of each pocket independently. A first grid was placed around the active site while a second was located around the DNP pocket. A low score was observed for the active site in both instances ($JEDI_{score} = 2.3$ and 2.6 respectively), because of a very high polarity caused by several ions buried in the active site, and the presence of numerous polar and charged amino acids. The DNP pocket in PSMA/8 does score slightly higher ($JEDI_{score} = 3.1$) than the same region in the PSMA/7 complex ($JEDI_{score} = 2.3$) but the score remains small because the DNP pocket is relatively small. Thus the PSMA binding site is a good illustration of challenging conditions encountered when performing JEDI analysis of binding sites for ligand that depart from typical rule-of-five compliant

small molecules.

HIV-RT is an enzyme playing a crucial role in the replication of the HIV virus. Several non-nucleoside RT inhibitors (NNRTIs) are already on the market.[45–49] Druggability predictions were compared for the NNRTI-binding pocket of the apo structure of HIV-1 RT and in complex with 9 (Figure 4.6E). This compound belongs to the second generation of NNRTIs and inhibits wild type HIV-RT with an IC50 of 2.1 nM (antiviral assay).[50, 51] The binding site of the holo system was found to be one of the most druggable pocket analyzed in this work. It is noteworthy that the NNRTI cavity is actually partially formed in the apo protein, and has an active volume of $V = 192\text{Å}^3$ The holo structure features an enlarged binding site and side chains rearrangements that increase the hydrophobicity Ha (Table 4.7, fifth row).

Interleukin-2 (IL-2) is a cytokine playing a crucial role in the regulation of white blood cells of the immune system. The small molecule 10 binds to a pocket only partially present in the apo structure. An additional cavity is present in the holo complex and it forms by displacement of two residues, Phe42 and Glu62 (Figure 4.6F).[52] A similar pocket volume descriptor is observed for both apo and holo form of IL-2. However this time, a higher druggability score was predicted in absence of ligand, because the hydrophobicity Ha is lower in the IL-2/10 complex (Table 4.7, sixth row). This occurred because the motion of Phe42 and Glu62 promotes hydrogen bonding with Glu62 and Lys43, activating grid points close to polar atoms, thus decreasing hydrophobicity.

Overall, the methodology is clearly able to correlate fluctuations in druggability score with noteworthy binding site conformational changes that have the potential to impact structure-based ligand design activities. In five cases out of six, the conformation with the highest $JEDI_{score}$ corresponds to the conformation

that binds the most tightly bound ligand. Careful interpretation of the results is needed when considering unusual protein-ligand complexes, such as PSMA/8. Quantitative correlation with binding affinities is not expected since the ligands differ. Further, druggability is not exclusively linked to binding affinity. PSMA is an example of a binding site for which ligands with very low K_i values are known (7 and 8), but the low predicted druggability score is adequate since most of the binding affinity is achieved by means of strongly polar ligand moieties positioned in the active site. These in turn translate into inauspicious drug-like properties such as low cell permeability.[42, 44]

4.4 Conclusion

A novel approach to assess protein binding site druggability has been developed. The fast, continuous and twice differentiable JEDI druggability estimator has been implemented in PLUMED and has been used as a collective variable in order to compute protein druggability at every integration step of a MD simulation.[24] The methodology is able to distinguish nondruggable, difficult and druggable pockets ($r^2 = 0.63$), and is relatively insensitive to insignificant structural rearrangements in a binding site. Some limits in the estimator were exposed, for instance neglect of potential metal-ligand interactions. This could be remedied with additional structural descriptors. JEDI was tested additionally on a dataset of hidden pockets for structurally diverse protein targets. The results show a good ability for the approach to detect structural modifications that influence the druggability of a protein binding site. With the present version of the method, care must be taken when performing this analysis on binding sites

for ligands that depart from typical rule-of-five compliant small molecules.

4.4 Bibliography

[1] Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, 3(8):711–715, August 2004.

[2] David Brown and Giulio Superti-Furga. Rediscovering the sweet spot in drug discovery. *Drug Discovery Today*, 8(23):1067–1077, December 2003.

[3] Andrew L Hopkins and Colin R Groom. The druggable genome. *Nature Reviews Drug Discovery*, 1(1):727–730, September 2002.

[4] Andreas P Russ and Stefan Lampel. The druggable genome: an update. *Drug Discovery Today*, 10(23-24):1607–1610, December 2005.

[5] Paul G Wyatt, Ian H Gilbert, Kevin D Read, and Alan H Fairlamb. Target validation: linking target and chemical properties to desired product profile. *Current Topics in Medicinal Chemistry*, 11(10):1275–1283, May 2011.

[6] Robert P Sheridan, Vladimir N Maiorov, M Katharine Holloway, Wendy D Cornell, and Ying-Duo Gao. Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *The Journal of Chemical Physics*, 50(11):2029–2040, November 2010.

[7] Fredrik N B Edfeldt, Rutger H A Folmer, and Alexander L Breeze. Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discovery Today*, 16(7-8):284–287, April 2011.

[8] Peter J Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry*, 28(7):849–857, July 1985.

[9] Frank Guarnieri and Mihaly Mezei. Simulated Annealing of Chemical Potential: A General Procedure for Locating Bound Waters. Application to the Study of the Differential Hydration Propensities of the Major and Minor

Grooves of DNA. *Journal of the American Chemical Society*, 118(35):8493–8494, January 1996.

[10] Collin M Stultz and Martin Karplus. MCSS functionality maps for a flexible protein. *Proteins*, 37(4):512–529, December 1999.

[11] Sheldon Dennis, Carlos J Camacho, and Sandor Vajda. Continuum electrostatic analysis of preferred solvation sites around proteins in solution. *Proteins*, 38(2):176–188, February 2000.

[12] Philip J Hajduk, Jeffrey R Huth, and Stephen W Fesik. Druggability indices for protein targets derived from NMR-based screening data. *Journal of Medicinal Chemistry*, 48(7):2518–2525, April 2005.

[13] Alan C Cheng, Ryan G Coleman, Kathleen T Smyth, Qing Cao, Patricia Soulard, Daniel R Caffrey, Anna C Salzberg, and Enoch S Huang. Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnology*, 25(1):71–75, January 2007.

[14] Tom Halgren. New method for fast and accurate binding-site identification and analysis. *Chemical Biology & Drug Design*, 69(2):146–148, February 2007.

[15] Thomas A Halgren. Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling*, 49(2):377–389, February 2009.

[16] Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, 10:168–168, January 2009.

[17] Peter Schmidtke and Xavier Barril. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *Journal of Medicinal Chemistry*, 53(15):5858–5867, August 2010.

[18] Agata Krasowski, Daniel Muthas, Aurijit Sarkar, Stefan Schmitt, and Ruth Brenk. DrugPred: a structure-based approach to predict protein druggability developed using an extensive nonredundant data set. *Journal of Chemical Information and Modeling*, 51(11):2829–2842, November 2011.

[19] Pietro Cozzini, Glen E Kellogg, Francesca Spyrakis, Donald J Abraham, Gabriele Costantino, Andrew Emerson, Francesca Fanelli, Holger Gohlke, Leslie A Kuhn, Garrett M Morris, Modesto Orozco, Thelma A Pertinhez, Menico Rizzi, and Christoph A Sotriffer. Target flexibility: an emerging consideration in drug discovery and design. *Journal of Medicinal Chemistry*, 51(20):6237–6255, October 2008.

[20] Inge Van Molle, Andreas Thomann, Dennis L Buckley, Ernest C So, Steffen Lang, Craig M Crews, and Alessio Ciulli. Dissecting fragment-based lead discovery at the von Hippel-Lindau Protein: Hypoxia Inducible Factor 1alpha protein-protein interface. *Chemistry & Biology*, 19(10):1300–1312, October 2012.

[21] Jesus Seco, F Javier Luque, and Xavier Barril. Binding site detection and druggability index from first principles. *Journal of Medicinal Chemistry*, 52(8):2363–2371, April 2009.

[22] Katrina W Lexa and Heather A Carlson. Full protein flexibility is essential for proper hot-spot mapping. *Journal of the American Chemical Society*, 133(2):200–202, January 2011.

[23] Ahmet Bakan, Neysa Nevins, Ami S Lakdawala, and Ivet Bahar. Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. *Journal of Chemical Theory and Computation*, 8(7):2435–2447, July 2012.

[24] Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provasi, Paolo Raiteri, Davide Donadio, Fabrizio Marinelli, Fabio Pietrucci, Ricardo A Broglia, and Michele Parrinello. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications*, 180(10):1961–1972, October 2009.

[25] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, September 1976.

[26] Andrew Butterfield, Vivek Vedagiri, Edward Lang, Cath Lawrence, Matthew J Wakefield, Alexander Isaev, and Gavin A Huttley. PyEvolve: a toolkit for statistical modelling of molecular evolution. *BMC Bioinformatics*, 5:1, January 2004.

[27] Andrea Volkamer, Daniel Kuhn, Friedrich Rippmann, and Matthias Rarey. DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics (Oxford, England)*, 28(15):2074–2075, August 2012.

[28] Emanuele Perola, Lee Herman, and Jonathan Weiss. Development of a rule-based method for the assessment of protein druggability. *Journal of Chemical Information and Modeling*, 52(4):1027–1038, April 2012.

[29] Jianghong An, Maxim Totrov, and Ruben Abagyan. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular & Cellular Proteomics : MCP*, 4(6):752–761, June 2005.

[30] Susanne Eyrisch and Volkhard Helms. Transient pockets on protein surfaces involved in protein-protein interaction. *Journal of Medicinal Chemistry*, 50(15):3457–3464, July 2007.

[31] Stéphanie Pérot, Olivier Sperandio, Maria A Miteva, Anne-Claude Camproux, and Bruno O Villoutreix. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today*, 15(15-16):656–667, August 2010.

[32] Douglas G Harris, Douglas P Marx, Jonathan M Anderson, Ronald W McCune, and S Scott Zimmerman. Kinetic and molecular modeling of nucleoside and nucleotide inhibition of malate dehydrogenase. *Nucleosides, Nucleotides & Nucleic Acids*, 21(11-12):813–823, November 2002.

[33] Kazuo Yamada, Nobumasa Hara, Tomoko Shibata, Harumi Osago, and Mikako Tsuchiya. The simultaneous measurement of nicotinamide adenine dinucleotide and related compounds by liquid chromatography/electrospray ionization tandem mass spectrometry. *Analitycal Biochemistry*, 352(2):282–285, May 2006.

[34] Gregory J Miller, Monita P Wilson, Philip W Majerus, and James H Hurley. Specificity determinants in inositol polyphosphate synthesis: crystal structure of inositol 1,3,4-trisphosphate 5/6-kinase. *Molecular Cell*, 18(2):201–212, April 2005.

[35] Jessica Huyet, Grégory M Pinon, Michel R Fay, Jérôme Fagart, and Marie-Edith Rafestin-Oblin. Structural basis of spirolactone recognition by the

mineralocorticoid receptor. *Molecular Pharmacology*, 72(3):563–571, September 2007.

[36] Deborah M Roll, Laurel R Barbieri, Ramunas Bigelis, Leonard A McDonald, Daniel A Arias, Li-Ping Chang, Maya P Singh, Scott W Luckman, Thomas J Berrodin, and Matthew R Yudt. The lecanindoles, nonsteroidal progestins from the terrestrial fungus Verticillium lecanii 6144. *Journal of Natural Products*, 72(11):1944–1948, November 2009.

[37] Anna Di Fiore, Carlo Pedone, Jochen Antel, Harald Waldeck, Andreas Witte, Michael Wurl, Andrea Scozzafava, Claudiu T Supuran, and Giuseppina De Simone. Carbonic anhydrase inhibitors: the X-ray crystal structure of ethoxzolamide complexed to human isoform II reveals the importance of thr200 and gln92 for obtaining tight-binding inhibitors. *Bioorganic & Medicinal Chemistry Letters*, 18(8):2669–2674, April 2008.

[38] Gary L Grunewald, Mitchell R Seim, Rachel C Regier, and Kevin R Criscione. Exploring the active site of phenylethanolamine N-methyltransferase with 1,2,3,4-tetrahydrobenz[h]isoquinoline inhibitors. *Bioorganic & Medicinal Chemistry*, 15(3):1298–1310, February 2007.

[39] Matthew E Cockman, Norma Masson, David R Mole, Panu Jaakkola, Gin-Wen Chang, Steven C Clifford, Eamonn R Maher, Christofer W Pugh, Peter J Ratcliffe, and Patrick H Maxwell. Hypoxia inducible factor-alpha binding and ubiquitylation by the von Hippel-Lindau tumor suppressor protein. *Journal of Biological Chemistry*, 275(33):25733–25741, August 2000.

[40] Michael Kothe, Darcy Kohls, Simon Low, Rocco Coli, Alan C Cheng, Suzanne L Jacques, Theresa L Johnson, Cristina Lewis, Christine Loh, Jim Nonomiya, Alissa L Sheils, Kimberly A Verdries, Thomas A Wynn, Cyrille Kuhn, and Yuan-Hua Ding. Structure of the catalytic domain of human polo-like kinase 1. *Biochemistry*, 46(20):5960–5971, May 2007.

[41] Robert A Elling, Raymond V Fucini, Emily J Hanan, Kenneth J Barr, Jiang Zhu, Kumar Paulvannan, Wenjin Yang, and Michael J Romanowski. Structure of the Brachydanio rerio Polo-like kinase 1 (Plk1) catalytic domain in complex with an extended inhibitor targeting the adaptive pocket of the enzyme. *Acta crystallographica. Section F, Structural Biology and Crystallization Communications*, 64(Pt 8):686–691, August 2008.

[42] Haofan Wang, Youngjoo Byun, Cyril Barinka, Mrudula Pullambhatla, Hyoeun C Bhang, James J Fox, Jacek Lubkowski, Ronnie C Mease, and Martin G Pomper. Bioisosterism of urea-based GCPII inhibitors: Synthesis and structure–activity relationship studies. *Bioorganic & Medicinal Chemistry Letters*, 20(1):392–397, January 2010.

[43] Andrew X Zhang, Ryan P Murelli, Cyril Barinka, Julien Michel, Alexandra Cocleaza, William L Jorgensen, Jacek Lubkowski, and David A Spiegel. A remote arene-binding site on prostate specific membrane antigen revealed by antibody-recruiting small molecules. *Journal of the American Chemical Society*, 132(36):12711–12716, September 2010.

[44] Ryan P Murelli, Andrew X Zhang, Julien Michel, William L Jorgensen, and David A Spiegel. Chemical control over immune recognition: a class of antibody-recruiting small molecules that target prostate cancer. *Journal of the American Chemical Society*, 131(47):17090–17092, December 2009.

[45] Peter M Grob, Joe C Wu, Kenneth A Cohen, Richard H Ingraham, Cheng-Kon Shih, Karl D Hargrave, Tari L McTague, and Vincent J Merluzzi. Nonnucleoside inhibitors of HIV-1 reverse transcriptase: nevirapine as a prototype drug. *AIDS Research and Human Retroviruses*, 8(2):145–152, February 1992.

[46] Donna L Romero, Raymond A Morge, Michael J Genin, Carolyn Biles, Mariano Busso, Lionel Resnick, Irene W Althaus, Fritz Reusser, Richard C Thomas, and William G Tarpley. Bis(heteroaryl)piperazine (BHAP) reverse transcriptase inhibitors: structure-activity relationships of novel substituted indole analogs and the identification of 1-[(5-methanesulfonamido-1H-indol-2-yl)carbonyl]-4-[3-[(1-methylethyl)amino]pyridinyl]piperazinemonomethanesulfonate (U-90152S), a second-generation clinical candidate. *Journal of Medicinal Chemistry*, 36(10):1505–1508, May 1993.

[47] Donna L Romero, Robert A Olmsted, Toni Jo Poel, Raymond A Morge, Carolyn Biles, Barbara J Keiser, Laurice A Kopta, Jan M Friis, John D Hosley, Kevin J Stefanski, Donn G Wishka, David B Evans, Joel Morris, Randy G Stehle, Satish K Sharma, Yoshihiko Yagi, Richard L Voorman, Wade J Adams, W Gary Tarpley, and Richard C Thomas. Targeting Delavirdine/Atevirdine Resistant HIV-1: Identification of (Alkylamino)piperidine-

Containing Bis(heteroaryl)piperazines as Broad Spectrum HIV-1 Reverse Transcriptase Inhibitors. *Journal of Medicinal Chemistry*, 39(19):3769–3789, January 1996.

[48] Robert M Esnouf, Jingshan Ren, Andrew L Hopkins, Carl K Ross, E Yvonne Jones, David K Stammers, and David I Stuart. Unique features in the structure of the complex between HIV-1 reverse transcriptase and the bis(heteroaryl)piperazine (BHAP) U-90152 explain resistance mutations for this nonnucleoside inhibitor. *Proceedings of the National Academy of Sciences of the United States of America*, 94(8):3984–3989, January 1997.

[49] Steven D Young, Susan F Britcher, Lee O Tran, Linda S Payne, William C Lumma, Terry A Lyle, Joel R Huff, Paul S Anderson, David B Olsen, Steven S Carroll, Douglas J Pettibone, Julie A O'Brien, Richard G Ball, Suresh K Balani, Jiunn H Lin, I-WU Chen, William A Schleif, Vinod V Sardana, William J Long, Vera W Byrnes, and Emilio A Emini. L-743, 726 (DMP-266): a novel, highly potent nonnucleoside inhibitor of the human immunodeficiency virus type 1 reverse transcriptase. *Antimicrobial Agents and Chemotherapy*, 39(12):2602–2605, December 1995.

[50] Yu Hsiou, Jianping Ding, Kalyan Das, Arthur D Clark, Steven H Hughes, and Edward Arnold. Structure of unliganded HIV-1 reverse transcriptase at 2.7 A resolution: implications of conformational changes for polymerization and inhibition mechanisms. *Structure*, 4(7):853–860, July 1996.

[51] Denis J Kertesz, Christine Brotherton-Pleiss, Minmin Yang, Zhanguo Wang, Xianfeng Lin, Zongxing Qiu, Donald R Hirschfeld, Shelley Gleason, Taraneh Mirzadegan, Pete W Dunten, Seth F Harris, Armando G Villaseñor, Julie Qi Hang, Gabrielle M Heilek, and Klaus Klumpp. Discovery of piperidin-4-yl-aminopyrimidines as HIV-1 reverse transcriptase inhibitors. N-Benzyl derivatives with broad potency against resistant mutant viruses. *Bioorganic & Medicinal Chemistry Letters*, 20(14):4215–4218, July 2010.

[52] Michelle R Arkin, Mike Randal, Warren L DeLano, Jennifer Hyde, Tinh N Luong, Johan D Oslob, Darren R Raphael, Lisa Taylor, Jun Wang, Robert S McDowell, James A Wells, and Andrew C Braisted. Binding of small molecules to an adaptive protein-protein interface. *Proceedings of the National Academy of Sciences of the United States of America*, 100(4):1603–1608, February 2003.

5

JEDI: derivatives and dynamics

This chapter covers the different results obtained using the JEDI approach during classical and biased molecular dynamics simulations

5.1 Introduction

The previous chapter introduced a new collective variable (CV) called JEDI algorithm ('Just Exploring Druggability at protein Interfaces'). JEDI has been designed to evaluate protein druggability 'on-the-fly' during molecular dynamics (MD) simulations without any organic probes or protein restraints. The druggability function relies on a set of geometric parameters describing the volume, the enclosure and the hydrophobicity of a binding site. Previously, the ability of JEDI to predict druggability fluctuations during classical MD simulations has been discussed.

The main novelty of the approach is that the JEDI scoring function is fast, continuous and differentiable. Accordingly, it can be used as a CV to bias MD simulations and enhance sampling of protein conformations. JEDI has been implemented in the software PLUMED 1.3 to enable metadynamics simulations and free-energy calculations with the most popular MD engines.[1] The methodology was parameterized using the freely accessible Druggable Cavity Directory (DCD) dataset.[2] This chapter aims to evaluate the potential for JEDI analyses to detect cryptic druggable binding sites in proteins. Two different systems, VHL and hPNMT, were selected to perform several umbrella sampling simulations. The first system has a binding site exposed to the solvent while it is buried for the second. However, both are known to adopt local structural rearrangements that influence the protein druggability.

5.2 Materials & Methods

5.2.1 JEDI derivatives

As described in the previous chapter, the JEDI potential is made of a combination of two structural descriptors (eq 5.1).

$$JEDI_{score} = V_{druglike}\left(\alpha V_a + \beta H_a + \gamma\right) \tag{5.1}$$

Because the JEDI potential is based on functions that are continuous and differentiable, the gradient with respect to the Cartesian coordinates x, y, z of the protein atoms can be calculated using the following equation:

$$\nabla JEDI_{score} = \sum_{j=1}^{M}\left(\frac{\partial JEDI_{score}}{\partial x_{p_j}} + \frac{\partial JEDI_{score}}{\partial y_{p_j}} + \frac{\partial JEDI_{score}}{\partial z_{p_j}}\right) \tag{5.2}$$

Where M is the number of protein atoms in the binding site region, $\frac{\partial JEDI_{score}}{\partial x_{p_j}}$, $\frac{\partial JEDI_{score}}{\partial y_{p_j}}$ and $\frac{\partial JEDI_{score}}{\partial z_{p_j}}$ are the partial derivatives with respect to the Cartesian coordinates of protein atom j. The derivative of the JEDI potential with respect to grid Cartesian coordinates does not need to be calculated as the grid is frozen during MD time-steps. By application of the product rule:

$$\frac{\partial JEDI_{score}}{\partial x_{p_j}} = JEDI_{score}\left[\frac{1}{V_{druglike}}\frac{\partial V_{druglike}}{\partial x_{p_j}} + \frac{1}{\alpha V_a + \beta H_a + \gamma}\left(\alpha\frac{\partial V_a}{\partial x_{p_j}} + \beta\frac{\partial H_a}{\partial x_{p_j}}\right)\right] \tag{5.3}$$

Similar equations can be derived for the two other partial derivatives with respect to y and z Cartesian coordinates.

In the context of a quadratic function (e.g. umbrella sampling), the JEDI potential is calculated as follows:

$$U_{JEDI} = k\left(JEDI_{score} - JEDI_{score}^{target}\right)^2 \tag{5.4}$$

The partial derivatives are:

$$\frac{\partial U_{JEDI}}{\partial x_{p_j}} = 2k\frac{\partial JEDI_{score}}{\partial x_{p_j}}\left(JEDI_{score} - JEDI_{score}^{target}\right) \tag{5.5}$$

5.2.1.1 Switching function

All the descriptors presented below are based on cubic splines such that the JEDI potential is continuous and twice differentiable. Two forms of switching functions have been used operating on variables v and k. The first one turns off with v starting at k at v_{min}, reaching 0 at $v_{min} + \Delta$ (5.6).

$$S_v^{off}\left(k, v_{min}, \Delta\right) = \begin{cases} k \text{ if } m < 0 \\ k\left[\left(1-m^2\right)^2\left(1+2m^2\right)\right] \text{ if } 0 \leq m \leq 1 \\ 0 \text{ if } m > 1 \end{cases} \tag{5.6}$$

where $m = \frac{v - v_{min}}{\Delta}$.

The partial derivatives are:

$$\frac{\partial S_v^{off}}{\partial x_{p_j}} = \frac{\partial S_v^{off}}{\partial m}\frac{\partial m}{\partial v}\frac{\partial v}{\partial x_{p_j}} + \frac{\partial S_v^{off}}{\partial k}\frac{\partial k}{\partial x_{p_j}} \tag{5.7}$$

Where

$$\frac{\partial S_v^{off}}{\partial m} = \begin{cases} 0 \text{ if } m < 0 \\ 4km\left[(1-m^2)^2 - (1-m^2)(1+2m^2)\right] \text{ if } 0 \leq m \leq 1 \\ 0 \text{ if } m > 1 \end{cases} \tag{5.8}$$

$$\frac{\partial m}{\partial v} = \frac{1}{\Delta} \tag{5.9}$$

$$\frac{\partial S_v^{off}}{\partial k} = \begin{cases} 1 \text{ if } m < 0 \\ (1-m^2)^2(1+2m^2) \text{ if } 0 \leq m \leq 1 \\ 0 \text{ if } m > 1 \end{cases} \tag{5.10}$$

The second form turns 'on' the variable S from 0 to k along an interval Δ (Equtation 5.11).

$$S_v^{on}(k, v_{min}, \Delta) = \begin{cases} 0 \text{ if } m < 0 \\ k\left[1 - (1-m^2)^2(1+2m^2)\right] \text{ if } 0 \leq m \leq 1 \\ k \text{ if } m > 1 \end{cases} \tag{5.11}$$

The corresponding partial derivatives are defined as:

$$\frac{\partial S_v^{on}}{\partial x_{p_j}} = \frac{\partial S_v^{on}}{\partial m}\frac{\partial m}{\partial v}\frac{\partial v}{\partial x_{p_j}} + \frac{\partial S_v^{on}}{\partial k}\frac{\partial k}{\partial x_{p_j}} \tag{5.12}$$

Where

$$\frac{\partial S_v^{on}}{\partial m} = \begin{cases} 0 \text{ if } m < 0 \\ -4km\left[\left(1-m^2\right)^2 - \left(1-m^2\right)\left(1+2m^2\right)\right] \text{ if } 0 \leq m \leq 1 \\ 0 \text{ if } m > 1 \end{cases} \tag{5.13}$$

$$\frac{\partial S_v^{off}}{\partial k} = \begin{cases} 0 \text{ if } m < 0 \\ 1 - \left(1-m^2\right)^2\left(1+2m^2\right) \text{ if } 0 \leq m \leq 1 \\ 1 \text{ if } m > 1 \end{cases} \tag{5.14}$$

An illustration of the switching functions and their derivatives is given in Figure 5.1.

5.2.1.2 Grid point activity

At every step of the MD simulation, an activity function assigns a score a between 0 and 1 to a grid point i such that grid points that are too close or too far to protein atoms have an activity of 0 (eq 5.15). Switching intervals are used to gradually activate grid points, so partial activity values are possible.

$$a_i = S_{BS_i}^{off}\left(1.0, BS_i, \Delta BS\right) S_{mind_i}^{on}\left(1.0, CC_{mind}, \Delta CC\right) S_{exposure_i}^{on}\left(1.0, E_{min}, \Delta E\right) \tag{5.15}$$

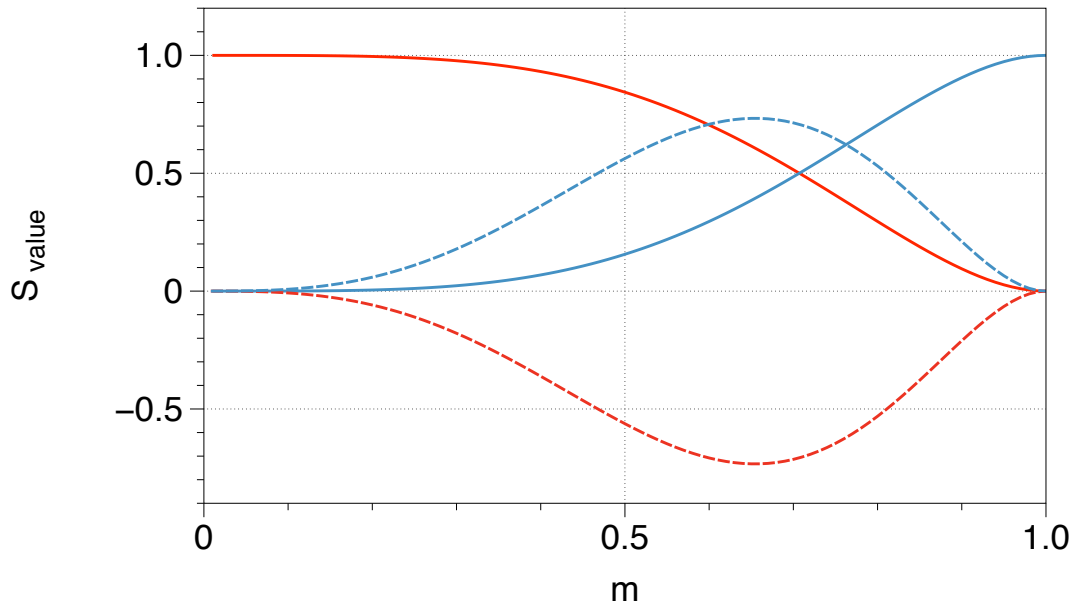

Figure 5.1: Representation of the two switching functions and first partial derivatives with respect to v, $k = 1$, v_{min}=0 and Δ=1. S_v^{on} and S_v^{off} are colored in blue and red respectively. The derivatives of each function with respect to m are represented as dashed lines.

The minimum distance between a grid point i and the M protein atoms is calculated as

$$BS_i = \frac{\theta}{\ln\left(\sum_{j=1}^{M} \exp\left(\frac{\theta}{\|\mathbf{r}_{ij}\|}\right)\right)} \tag{5.16}$$

With $\theta = 5$Å and $\mathbf{r}_{ij} = \mathbf{r}_{gi} - \mathbf{r}_{pj}$, where $\mathbf{r}_{gi}$ and $\mathbf{r}_{pj}$ are respectively the position vectors of grid point i and protein atom j belonging to the binding site region. The second term in equation 5.15 causes grid points that overlap with protein atoms to be gradually inactivated. The minimum distance $mind_i$ between grid points and protein atoms is calculated with an equation similar to eq 5.16. The third term in equation 5.15 gradually inactivates solvent exposed grid points.

$$exposure_i = \sum_{k=1}^{N} [S^{off}_{mind_k}(1.0, CC2_{min}, \Delta CC2)\, S^{on}_{\|\mathbf{r}_{ik}\|}(1.0, GP1_{min}, \Delta GP1) \\ S^{off}_{\|\mathbf{r}_{ik}\|}(1.0, GP2_{min}, \Delta GP2)] \tag{5.17}$$

where $CC2_{min}/\Delta CC2$ control the distance below which a grid point is considered as interacting with the protein. $GP1_{min}/\Delta GP1$ and $GP2_{min}/\Delta GP2$ are used to select grid points at a given distance interval from the grid point i in order to penalize solvent exposed grid points.

The partial derivatives of equation 5.15 are defined as:

$$\frac{\partial a_i}{\partial x_{p_j}} = \frac{\partial S^{off}_{BS_i}}{\partial x_{p_j}} S^{on}_{mind_i} S^{on}_{exposure_i} + \frac{\partial S^{on}_{mind_i}}{\partial x_{p_j}} S^{off}_{BS_i} S^{on}_{exposure_i} + \frac{\partial S^{on}_{exposure_i}}{\partial x_{p_j}} S^{off}_{BS_i} S^{on}_{mind_i} \tag{5.18}$$

Because the first term of the equation 13 does not vary with the atomic coordinate changes during a simulation, the partial derivative with respect to x can be simplified as follows:

$$\frac{\partial a_i}{\partial x_{p_j}} = S^{off}_{BS_i} \left(\frac{\partial S^{on}_{mind_i}}{\partial x_{p_j}} S^{on}_{exposure_i} + \frac{\partial S^{on}_{exposure_i}}{\partial x_{p_j}} S^{on}_{mind_i} \right) \tag{5.19}$$

where

$$\frac{\partial S^{on}_{mind_i}}{\partial x_{p_j}} = \frac{\partial S^{on}_{mind_i}}{\partial m} \frac{1}{\Delta CC} \frac{\partial mind_i}{\partial x_{p_j}} + \frac{\partial S^{on}_{mind_i}}{\partial k} \frac{\partial k}{\partial x_{p_j}} \tag{5.20}$$

Because k is a constant, the partial derivatives with respect to x_{p_j} are null. Consequently, the previous equation can be simplified. Only the following equations are needed:

$$\frac{\partial mind_i}{\partial x_{p_j}} = \frac{\partial mind_i}{\partial \|\mathbf{r}_{ij}\|} \frac{\partial \|\mathbf{r}_{ij}\|}{\partial x_{p_j}} \tag{5.21}$$

$$\frac{\partial mind_i}{\partial \|\mathbf{r}_{ij}\|} = \frac{\theta^2 e^{\frac{\theta}{\|\mathbf{r}_{ij}\|}}}{\|\mathbf{r}_{ij}\|^2 \sum_{j=1}^{M} \exp\left(\frac{\theta}{\|\mathbf{r}_{ij}\|}\right) \left(\ln \sum_{j=1}^{M} \exp\left(\frac{\theta}{\|\mathbf{r}_{ij}\|}\right)\right)^2} \tag{5.22}$$

$$\frac{\partial \|\mathbf{r}_{ij}\|}{\partial x_{p_j}} = -\frac{x_{g_i} - x_{p_j}}{\sqrt{\left(x_{g_i} - x_{p_j}\right)^2 + \left(y_{g_i} - y_{p_j}\right)^2 + \left(z_{g_i} - z_{p_j}\right)^2}} \tag{5.23}$$

And similarly:

$$\frac{\partial S^{on}_{exposure_i}}{\partial x_{p_j}} = \frac{\partial S^{on}_{exposure_i}}{\partial m} \frac{1}{\Delta E} \frac{\partial E_i}{\partial x_{p_j}} \tag{5.24}$$

Because grid points coordinates do not change during the simulation, $S^{on}_{\|\mathbf{r}_{ik}\|}\left(1.0, GP1_{min}, \Delta GP1\right)$ and $S^{off}_{\|\mathbf{r}_{ik}\|}\left(1.0, GP2_{min}, \Delta GP2\right)$ are considered as a constant. Thus, $\frac{\partial E_i}{\partial x_{p_j}}$ can be simplified and expressed as follows

$$\frac{\partial E_i}{\partial x_{p_j}} = \sum_{k=1}^{N} \left(\frac{\partial S^{off}_{mind_k}}{\partial m} \frac{1}{\Delta CC2} \frac{\partial mind_k}{\partial x_{p_j}} \right) S^{on}_{\|\mathbf{r}_{ik}\|}\left(1.0, GP1_{min}, \Delta GP1\right) S^{off}_{\|\mathbf{r}_{ik}\|}\left(1.0, GP2_{min}, \Delta GP2\right) \tag{5.25}$$

Where $\frac{\partial mind_k}{\partial x_{p_j}}$ is calculated as described in the equation 5.21.

5.2.1.3 Volume

The active volume descriptor (V) is calculated as the sum of the activity of the N grid points weighted by the volume of space V_g monitored by the grid point (1.5 $Å^3$ by default).

$$V = \sum_{i=1}^{N} a_i V_g \tag{5.26}$$

This volume is then divided by the maximum active volume descriptor (V_{max}) observed in the training dataset in order to obtain the pocket volume descriptor V_a.

$$V_a = \frac{V}{V_{max}} \tag{5.27}$$

The partial derivative with respect to x of the previous equation varies according to the activity a_i of each grid point and is calculated as

$$\frac{\partial V_a}{\partial x_{p_j}} = \frac{\partial V}{\partial x_{p_j}} \frac{1}{V_{max}} \tag{5.28}$$

where

$$\frac{\partial V}{\partial x_{p_j}} = \sum_{i=1}^{N} \frac{\partial a_i}{\partial x_{p_j}} V_g \tag{5.29}$$

The partial derivative of the activity is calculated as described in equation 5.18.

The active volume descriptor V is then converted into an overall switching factor ($V_{druglike}$) in the interval [0,1], with smaller values assigned to large active

volumes or small active volumes. The aim is to penalize the apparition of too large or too small cavities on the grid. The range of 'ideal' number of active grid points V_{max} to V_{min} has been defined according to both smaller and larger values in the training dataset such that no penalty applies for drug-like small molecules sized binding site.

$$V_{druglike} = S_v^{off}\left(1.0, V_{max}, \Delta V_{max}\right) S_v^{on}\left(1.0, V_{min}, \Delta V_{min}\right) \tag{5.30}$$

The partial derivative of $V_{druglike}$ with respect to x is computed as

$$\frac{\partial V_{druglike}}{\partial x_{p_j}} = S_v^{off}\frac{\partial V_v^{on}}{\partial x_{p_j}} + S_v^{on}\frac{\partial V_v^{off}}{\partial x_{p_j}} \tag{5.31}$$

where

$$\frac{\partial S_v^{on}}{\partial x_{p_j}} = \frac{\partial S_v^{on}}{\partial m}\frac{1}{\Delta V_{min}}\frac{\partial V}{\partial x_{p_j}} \tag{5.32}$$

and

$$\frac{\partial S_v^{off}}{\partial x_{p_j}} = \frac{\partial S_v^{off}}{\partial m}\frac{1}{\Delta V_{max}}\frac{\partial V}{\partial x_{p_j}} \tag{5.33}$$

5.2.1.4 Hydrophobicity

The active grid hydrophobicity function aims to capture the average hydrophobicity of the active grid points calculated as:

$$H_a = \frac{1}{V} \sum_{i=1}^{N} H_i a_i \tag{5.34}$$

with

$$H_i = \frac{apolar_i}{apolar_i + polar_i} \tag{5.35}$$

where $apolar_i$ and $polar_i$ are respectively the number of apolar and polar protein atoms within the distance r_{hydro}.

$$apolar_i = \sum_{j=1}^{M_{apolar}} S^{off}_{\|\mathbf{r}_{ij}\|}\left(a_i, r_{hydro}, \Delta r_{hydro}\right) \tag{5.36}$$

$$polar_i = \sum_{j=1}^{M_{polar}} S^{off}_{\|\mathbf{r}_{ij}\|}\left(a_i, r_{hydro}, \Delta r_{hydro}\right) \tag{5.37}$$

Partial derivatives are calculated as:

$$\frac{\partial H_a}{\partial x_{p_j}} = \frac{1}{V} \sum_{i=1}^{N} \left(\frac{\frac{\partial apolar_i}{\partial x_{p_j}} \left(apolar_i + polar_i\right) - apolar_i \left(\frac{\partial apolar_i}{\partial x_{p_j}} + \frac{\partial polar_i}{\partial x_{p_j}} \right)}{\left(apolar_i + polar_i\right)^2} \right) \tag{5.38}$$

However, the protein atom j cannot be both polar and apolar. Consequently, two different kinds of derivatives are calculated according to the polar or apolar property of the protein atom j.

If the protein atom j is polar, then $\frac{\partial apolar_i}{\partial x_{p_j}} = 0$

$$\frac{\partial H_i}{\partial x_{p_j}} = -\frac{apolar_i \frac{\partial polar_i}{\partial x_{p_j}}}{\left(apolar_i + polar_i\right)^2} \tag{5.39}$$

where

$$\frac{\partial polar_i}{\partial x_{p_j}} = \sum_{j=1}^{M_{polar}} \frac{\partial S^{off}_{\|\mathbf{r}_{ij}\|}}{\partial m} \frac{1}{\Delta r_{hydro}} \frac{\partial \|\mathbf{r}_{ij}\|}{\partial x_{p_j}} + \frac{\partial S^{off}_{\|\mathbf{r}_{ij}\|}}{\partial a_i} \frac{\partial a_i}{\partial x_{p_j}} \tag{5.40}$$

If the protein atom j is apolar, then $\frac{\partial apolar_i}{\partial x_{p_j}} = 0$

$$\frac{\partial H_i}{\partial x_{p_j}} = -\frac{\left(apolar_i + polar_i\right) \frac{\partial apolar_i}{\partial x_{p_j}} - apolar_i \frac{\partial apolar_i}{\partial x_{p_j}}}{\left(apolar_i + polar_i\right)^2} \tag{5.41}$$

where

$$\frac{\partial apolar_i}{\partial x_{p_j}} = \sum_{j=1}^{M_{apolar}} \frac{\partial S^{off}_{\|\mathbf{r}_{ij}\|}}{\partial m} \frac{1}{\Delta r_{hydro}} \frac{\partial \|\mathbf{r}_{ij}\|}{\partial x_{p_j}} + \frac{\partial S^{off}_{\|\mathbf{r}_{ij}\|}}{\partial a_i} \frac{\partial a_i}{\partial x_{p_j}} \tag{5.42}$$

5.2.2 Molecular Dynamics Simulations

Proteins, ligands and cofactors were prepared using the python script Protein Preparation Wizard developed by Schrodinger available in Maestro.[3] First, missing hydrogen atoms were added to the structure to assign the appropriate bond number and formal charge. Then, proteins were manually verified to avoid incomplete side chains and steric clashes. Molecular dynamics simulations have been performed using GROMACS 4.5.5 combined with PLUMED 1.3.[1, 4] Simulations were carried

out in implicit solvent using the Generalized Born model and the Onufriev-Bashford-Case method to calculate the Born radii with a cutoff 20 Å.[5, 6] An energy minimization was performed using the steepest descent algorithm to reach the convergence parameter of 300 kJ.mol^{-1}.nm^{-2} of maximum force change. Then, production runs of 50 ns were performed using a time step of 2.0 fs. Systems were maintained at a constant temperature of 310 K using a stochastic Berendsen thermostat with a coupling constant of 1.0 ps.[7] The force field Amber99sb-ILDN was used for the proteins and the GAFF force field has been used for ligands and cofactors.[8, 9] The GAFF parameters for the ligands and the cofactors were obtained by using the software acpype, in combination with the antechamber utility from the AMBER12 software package.[10, 11] A new non-charged atom type was created to represent grid points. To avoid interactions between the protein atoms and the grid points, the Lennard-Jones parameters σ and ε and the atomic partial charges were equal to zero. All grid points are frozen in space during energy minimization and molecular dynamics time-steps.

5.2.3 Umbrella Sampling Simulations

Several umbrella sampling calculations were performed using the following biasing potential:

$$V(s(\mathbf{r})) = \kappa(s(\mathbf{r}) - s_0)^2 \tag{5.43}$$

where $s(\mathbf{r})$ is the $JEDI_{score}$ of protein conformation $\mathbf{r}$, κ is the force constant of the biasing potential, and s_0 is a target value for $JEDI_{score}$.[12] Several biased MD simulations were performed by varying κ and s_0 for different systems. The

resulting trajectories were clustered to identify the most likely conformations associated with a given set of (κ, s_0) values. In order to identify the most representative conformations present in a trajectory, the single linkage clustering approach available in GROMACS was used. Two structures were considered as neighbor if the RMSD was inferior to 1 Å. RMSD calculations were performed using the coordinates of heavy atoms constituting the binding site, excluding atoms that can form symmetry equivalent conformations (e.g. Valine C_γ atoms). Finally, cluster homogeneity was manually checked.

5.2.4 Docking Calculations

Several representative protein structures were extracted from the trajectories to perform *in silico* docking experiments. First, hydrogen atoms from both receptors and ligands were removed using the software Maestro.[3] Then, the docking was realized using Autodock Vina and the Autodock/Vina plugin for pymol.[13, 14] For each complex, the same grid was used, and twenty scores and poses were estimated. To be consistent with the simulations, His110 and His115 were only protonated on the epsilon nitrogen.

Several representative protein structures were extracted from the trajectories to perform docking calculations. The Maestro software was used to prepare input files for both receptors and ligands.[3] For VHL, protonation states of binding site Histidine residues were chosen to be consistent with those from the MD simulations (in particular, His110 and His115 were protonated on the ε-nitrogen atom). Docking calculations were performed with the software Autodock Vina and the Autodock/Vina plugin for pymol.[13, 14] For each complex, the same

docking grid was used, and up to twenty poses were generated. Different protocols featuring a fully rigid receptor or allowing side-chain flexibility of selected residues were used.

5.3 Results

Two different systems from the hidden pocket dataset introduced in chapter 4 have been selected to perform classical MD simulations and umbrella sampling simulations using JEDI.

5.3.1 VHL

The von Hippel-Lindau protein (pVHL) interacts with CUL2, Elongin B and C, and Rbx1. This complex is involved in the ubiquitination of the transcription factor hypoxia-inducible factor (HIF) through pVHL, leading to proteasome-mediated degradation of HIF. Few small molecules were found to inhibit interactions between pVHL and HIF such as **1** or **2** with a respective K_d obtained by fluorescence polarization of 86.1 μM and 27.7 μM.15 Both ligands interact with the same binding site, however a different orientation of the Arg107 is observed (Figure 5.2). In this solvent exposed cavity example, our approach was also able to detect a more druggable conformation that shows increased Ha. However, the difference of druggability is limited by a small decrease of the active volume. Indeed, the displacement of Arg107 exposes more grid points to the

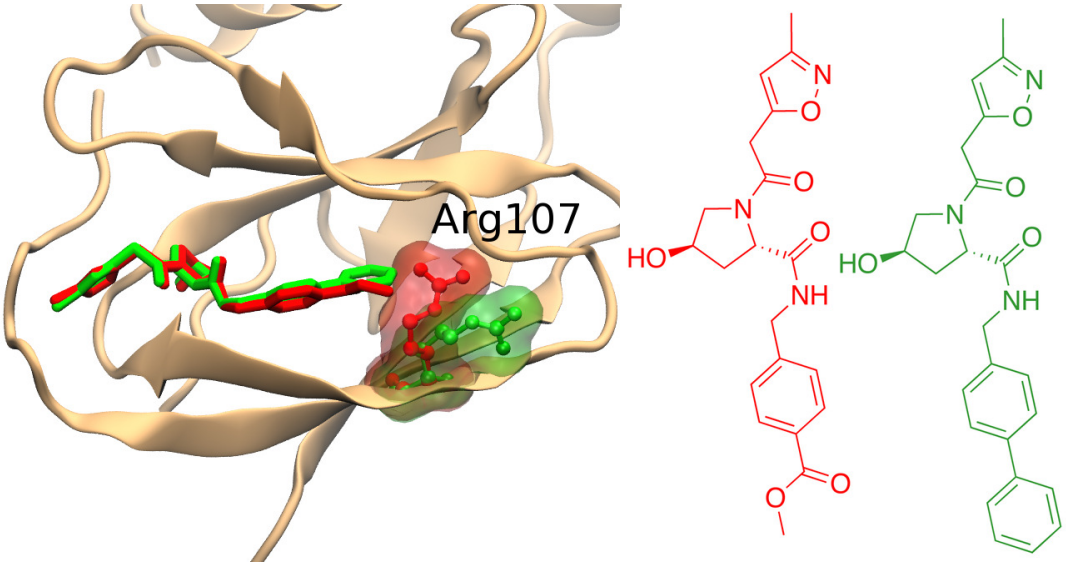

Figure 5.2: Illustration of the VHL binding site. The conformational changes inducing an increase of druggability (according to the literature) are highlighted in green (**2**) while the less druggable conformation is represented in red (**1**). The part of the binding site that is not involved in the druggability variation is colored brown. Ligand corresponding to each binding site is represented beside the figure. Pictures were prepared using the software VMD.

solvent. Consequently, grid points previously fully active become partially active (Table 5.1, first row).

Ligand	Protein	PDB code	$JEDI_{score}$	V / Å^3	H_a	$V_{druglike}$
1	VHL	3ZTD	8.2	118	0.80	1.0
2		3ZTC	8.5	114	0.82	1.0
3	hPNMT	1HNN	8.4	259	0.72	1.0
4		2G8N	9.2	276	0.74	1.0

Table 5.1: JEDI descriptor values for VHL and hPNMT.

5.3.1.1 Molecular Dynamics Simulations

Further tests were conducted with MD simulations of VHL. Druggability values were collected every ps over the course of a 50 ns simulation of apo VHL or VHL/**1**. The results are shown in Figure 5.3.

The binding site druggability remained stable throughout the VHL/**3** simulation, with an average $JEDI_{score}$ of 7.8 ± 0.6 which is consistent with the expected value from previous analyses (Table 5.1, first row). Clustering analysis with a RMSD cutoff of 1 Å reveals only one major binding site conformation (76% of the trajectory), that is depicted in Figure 5.3C (right panel). By contrast, the apo simulation shows an average druggability score of 5.7±0.8. Numerous structurally different binding site conformations are sampled. In the present MD simulations, the apo binding pocket is quickly obstructed by the rearrangement of Tyr98 and His110 inducing a drop of druggability. Dozens of clusters were identified and the most populated ($JEDI_{score}$ ca. 6.3) is present in 67% of the simulation (Figure 5.3C, left panel). This partially closed conformation is mainly stabilized by hydrogen bonds between the phenolic OH group of Tyr98 and the protein backbone. His110 is very flexible throughout the simulation. Surprisingly, significant side-chain rearrangements that partially block the binding site do not affect dramatically the $JEDI_{score}$ values. This occurs here because the shift in position for Tyr98 has created a new hydrophobic sub-pocket that contributes favorably to the $JEDI_{score}$. However this sub-pocket is now occluded by Tyr98 and disconnected from the rest of the binding site. Further, the rest of the VHL binding site is still partially present, including the central pyrrolidine binding pocket. Binding site conformations that correspond to extreme druggability fluctuations seen in the apo simulation are depicted in Figure 5.3D. In general, the apo

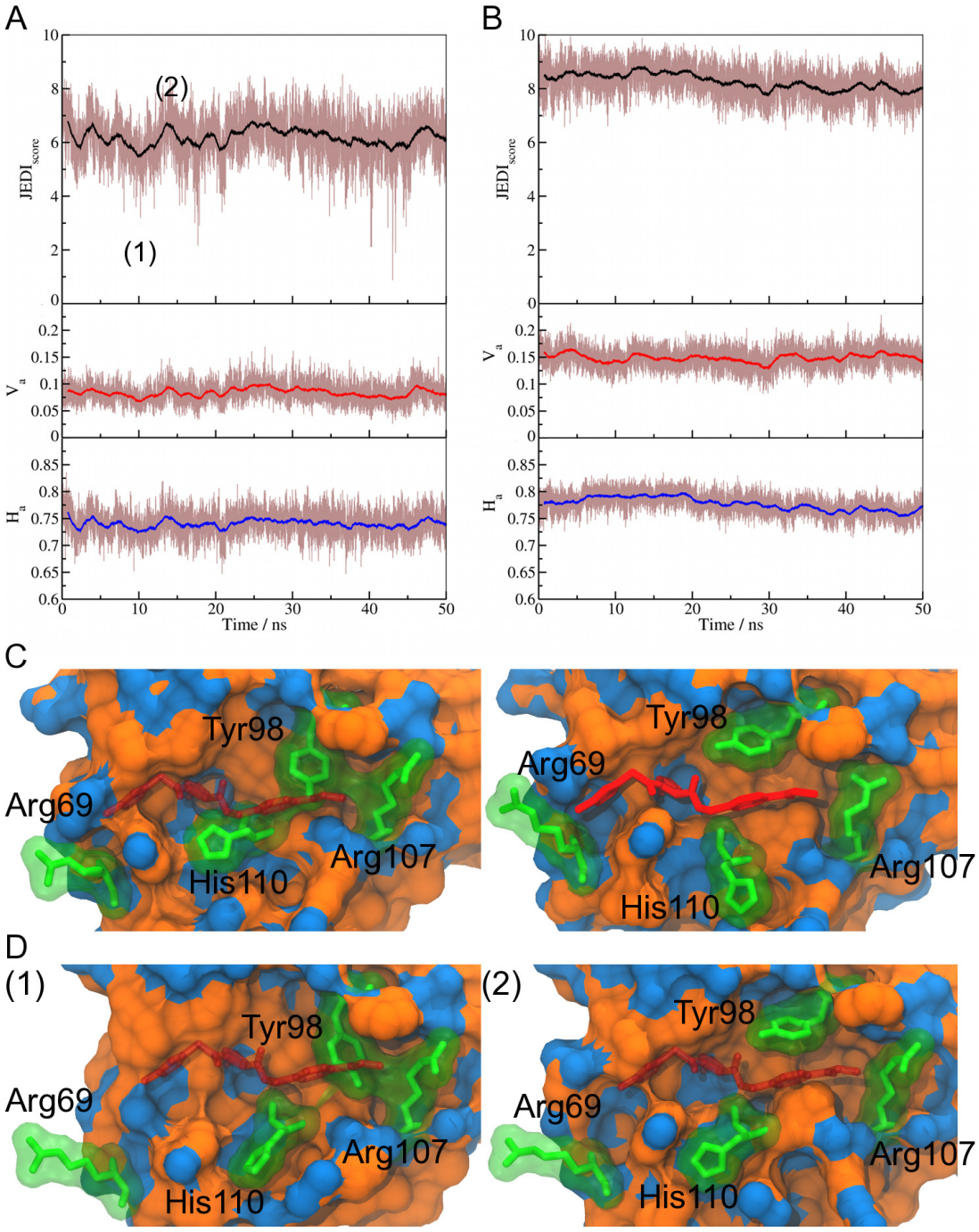

Figure 5.3: Druggability fluctuations during an MD simulations of apo VHL. Instantaneous values (thin lines) and 300 ps windowed averages (bold lines) of $JEDI_{score}$, Va and Ha during an MD simulation are represented in black, red and blue respectively for (A) apo VHL and (B) VHL/**1**. C) The most representative conformation of apo VHL (left) and VHL/**1** (right). D) Instantaneous conformations indicated by numbers 1 and 2 in panel A. Protein surface were colored according to polar (blue) and apolar atoms (orange). Protein residues discussed in the text are highlighted in green sticks. The ligand is represented in red sticks. The ligand (transparent red) was overlap with the conformations from apo VHL by structural alignment to indicate the position of the binding site. Pictures were prepared using the software VMD.

conformations that present high $JEDI_{score}$ values were found to be structurally very similar to the VHL/**1** conformation.

5.3.1.2 Umbrella Sampling Simulations

Umbrella sampling simulations were performed for apo VHL and VHL/**1** using equation 5.43 and by varying force constant values for κ and target $JEDI_{score}$ values s_0. The results are depicted in Figure 5.4. Apo simulations were biased to achieve a $JEDI_{score}$ of 8, in expectation with the values previously observed for ligand bound complexes (Table 5.1, first row). Figure 5.4A (left panel) indicates that the target druggability value is rapidly achieved in all instances. As expected fluctuations from the target value decrease with increased κ values. The trajectory obtained using $\kappa = 2000$ kJ.mol^{-1}.nm^{-2} was subjected to further clustering. The most populated clusters (51% of the overall trajectory) are very similar to the VHL/**1** structure, with RMSD values always inferior to 2.0 Å. In the unbiased MD simulation of apo VHL, only 14% of the computed conformation exhibited an RMSD to the VHL/**1** conformation that was smaller than 2.0 Å. Some clusters still contain conformations with Tyr98 pointing inside the binding site, but the occurrence is greatly decreased. His110 was also found to be much less flexible. It is apparent that the ligand binding site is almost fully formed in the most populated cluster of the biased apo VHL simulation (Figure 5.4A, right panel).

The umbrella sampling simulations of VHL/**1** were performed to encourage the binding site to adopt more druggable conformations. A reference value $s_0 = 9$ was selected based on the $JEDI_{score}$ of VHL/**2**. Figure 5.4B left panel shows that higher κ values are needed to achieve the desired s_0 value. This indicates that the conformations with high $JEDI_{score}$ values do not form spontaneously. The

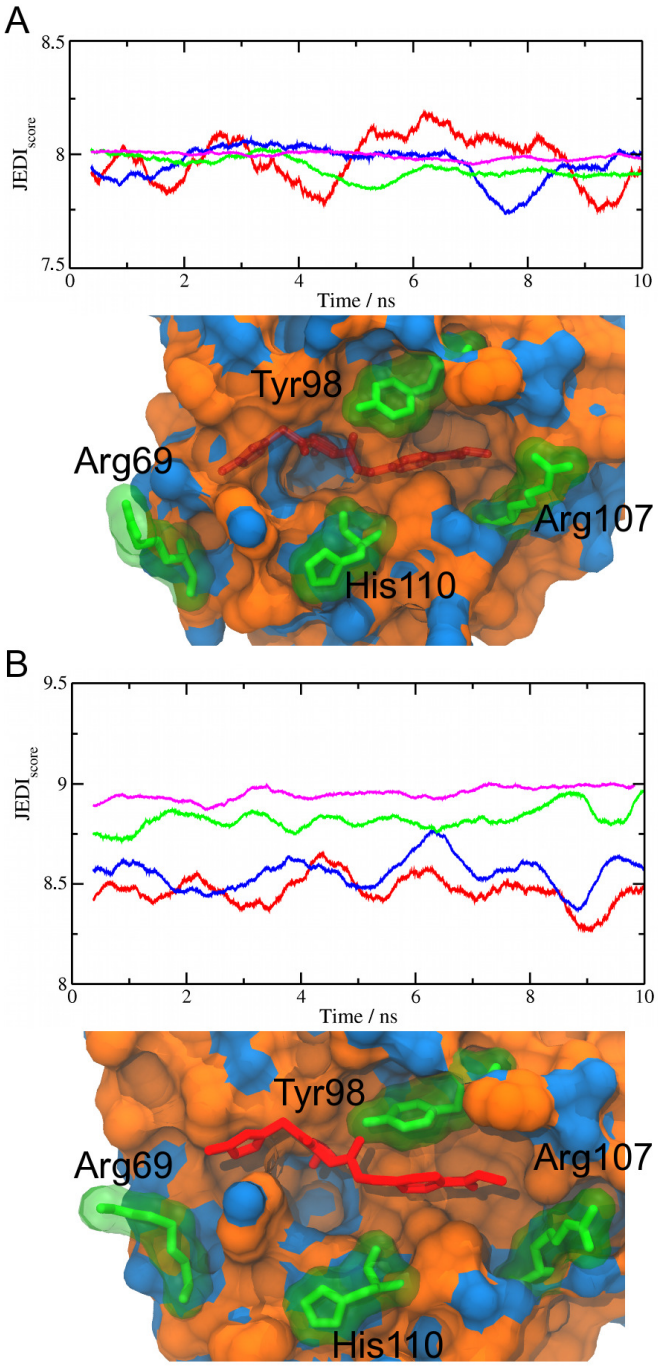

Figure 5.4: Druggability fluctuations during umbrella sampling simulations of (A) apo VHL and (B) VHL/**1**. For clarity, only the running averages are shown for four different spring constants (red: $\kappa = 500$ kJ.mol^{-1}.nm^{-2}, blue: $\kappa = 1000$ kJ.mol^{-1}.nm^{-2}, green: $\kappa = 2000$ kJ.mol^{-1}.nm^{-2}, magenta: $\kappa = 5000$ kJ.mol^{-1}.nm^{-2}). An illustration of the most populated cluster from the simulation performed with $\kappa = 2000$ kJ.mol^{-1}.nm^{-2} is depicted beside each graph. All other symbols and representations are as in Figure 5.3.

increase in $JEDI_{score}$ values that is achieved correlates largely with the position of Arg107. This amino acid initially closes the binding site, but with the present bias, it shifts rapidly to a solvent exposed position, thus causing an enlargement of the binding site. This motion was rarely observed in unbiased MD simulations.

Next, more significant structural rearrangements were sought by performing umbrella sampling simulations of apo VHL with $s_0 = 3.0$. Results obtained with $\kappa = 2000$ kJ.mol^{-1}.nm^{-2} are shown in Figure 5.5. Requesting such a low target druggability value forces VHL to largely collapse the binding site. Here the collapse is even more pronounced than observed in the unbiased apo VHL simulations, with the binding pockets of the isoxazole and pyrrolidine moieties completely masked. Consequently, the pocket volume descriptor V_a decreases, and the active volume V becomes sufficiently low such that the $V_{druglike}$ term penalizes the $JEDI_{score}$ values. The hydrophobicity descriptor H_a is stable during the biased simulation, with an average value slightly lower than observed in the unbiased apo VHL simulation. The closure of the binding site has totally or partially inactivated numerous grid points that were previously in a buried cavity, leaving only a few active grid points at the protein surface and near polar groups. An illustration of the most populated cluster (73% of the trajectory) is depicted in Figure 5.5B.

Umbrella sampling simulations of apo VHL were also performed by setting $s_0 = 10$ and $\kappa = 2000$ kJ.mol^{-1}.nm^{-2} to encourage the exploration of conformations with high druggability. The results are presented in Figure 5.6A. As observed previously, the simulation is rapidly sampling conformations in the requested range of $JEDI_{score}$. As expected, V_a and H_a are almost always higher than in the previously described simulations. However, larger fluctuations are observed in both descriptors throughout the biased simulation. An increase in

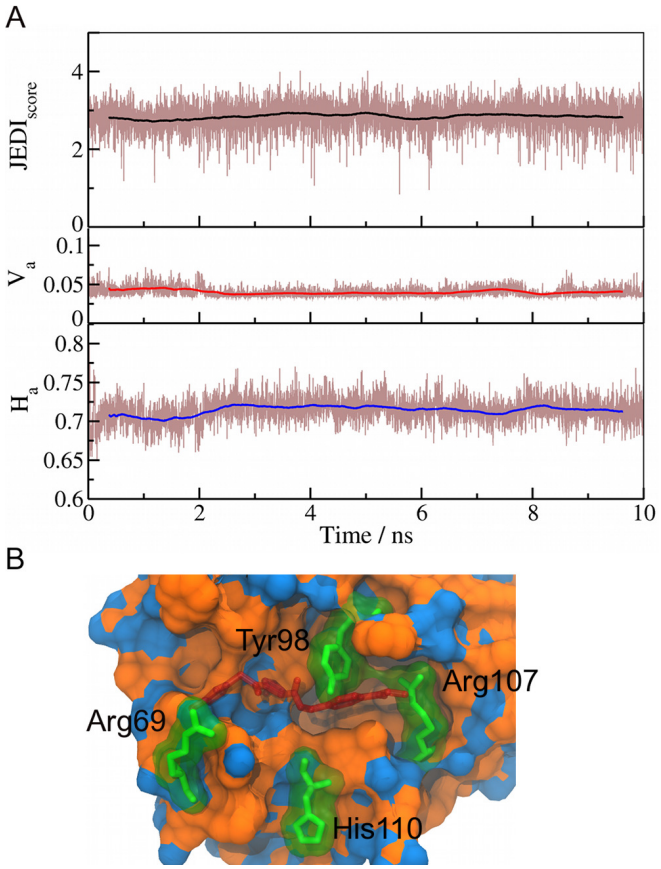

Figure 5.5: Druggability fluctuations during a biased simulation of apo VHL with $s_0 = 3$, $\kappa = 2000$ kJ.mol^{-1}.nm^{-2}. A) Instantaneous values and running averages of $JEDI_{score}$, V_a and H_a. B) Representative conformation of the most populated cluster identified in the simulation. All other symbols and representations are as in Figure 5.3.

hydrophobicity H_a is always offset by a decrease of the active volume descriptor V_a and vice versa. Clustering analysis of the trajectory here reveals at least two significant distinct clusters (populations 18% and 8% respectively). The second cluster (Figure 5.6C) corresponds to a low V_a / high H_a binding site conformation that is significantly different from the VHL/**1** structure. The pyrrolidine pocket has collapsed and side-chains rearranged to expose hydrophobic groups to the surface. The first cluster (Figure 5.6B) corresponds to a conformation comparable to the VHL/3 holo structure. Additionally, Arg107 has adopted a solvent exposed position that contributes favorably to the $JEDI_{score}$ as demonstrated previously (Table 5.1, second row). A significant difference that was not observed in previous simulations is the rearrangement of Arg69 in the left-hand side part of the binding site. This conformational rearrangement leads to a more extended cavity with high druggability scores. The flexibility of the left hand side pocket, has been recently discussed in the literature in the context of crystallographic structure analyses of multiple VHL ligand complexes,[15] and Galdeano et al. have suggested that additional interactions between ligands and this part of the binding site may facilitate the development of improved VHL ligands.

5.3.1.3 Docking

Several docking experiments were carried out to evaluate the conformational ensembles computed from the umbrella sampling simulations. Figure 5.7A depicts results obtained using the computed apo VHL conformation closest to the average conformation of the most populated cluster taken from an umbrella sampling simulation with $s_0 = 10.0$ and $\kappa = 2000$ kJ.mol^{-1}.nm^{-2} (Figure 5.7B, top).

Ligand **1** was found to adopt a pose that bears a substantial similarity

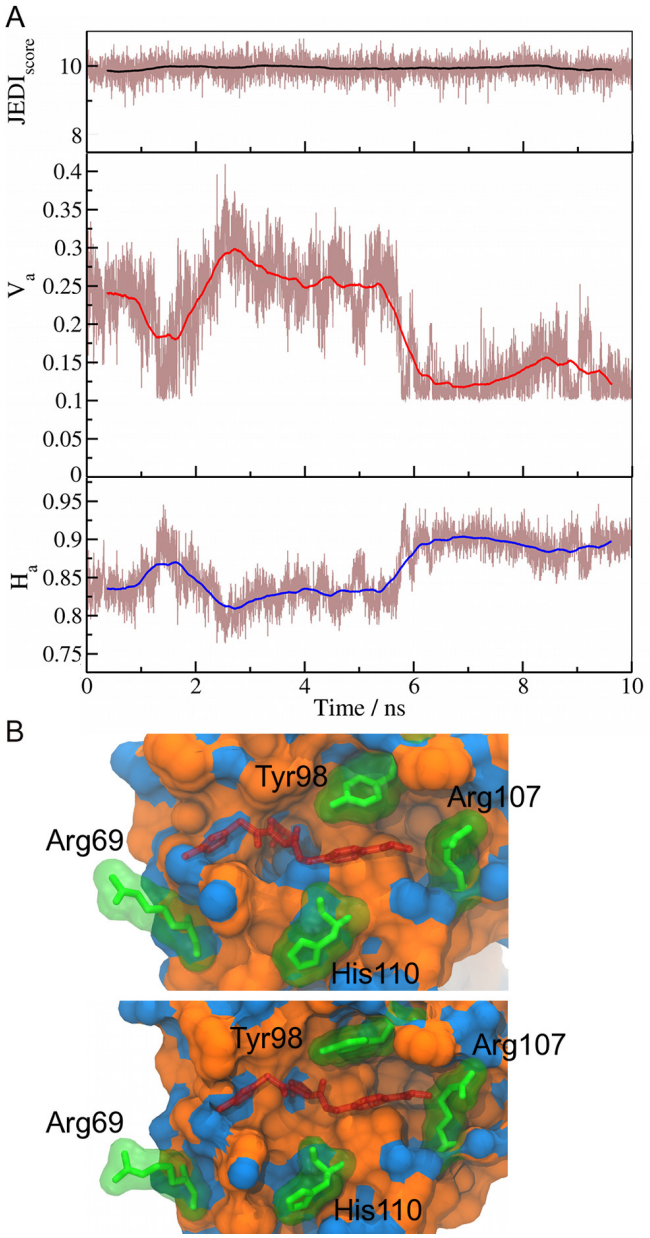

Figure 5.6: Druggability fluctuations in apo VHL umbrella sampling simulation with $s_0 = 10$ and $\kappa = 2000$ kJ.mol^{-1}.nm^{-2}. A) Running averages and instantaneous values of of $JEDI_{score}$, V_a and H_a, B) The most representative conformation of the first (top) and second (bottom) most populated clusters observed during the simulation. All other symbols and representations are as in Figure 5.3.

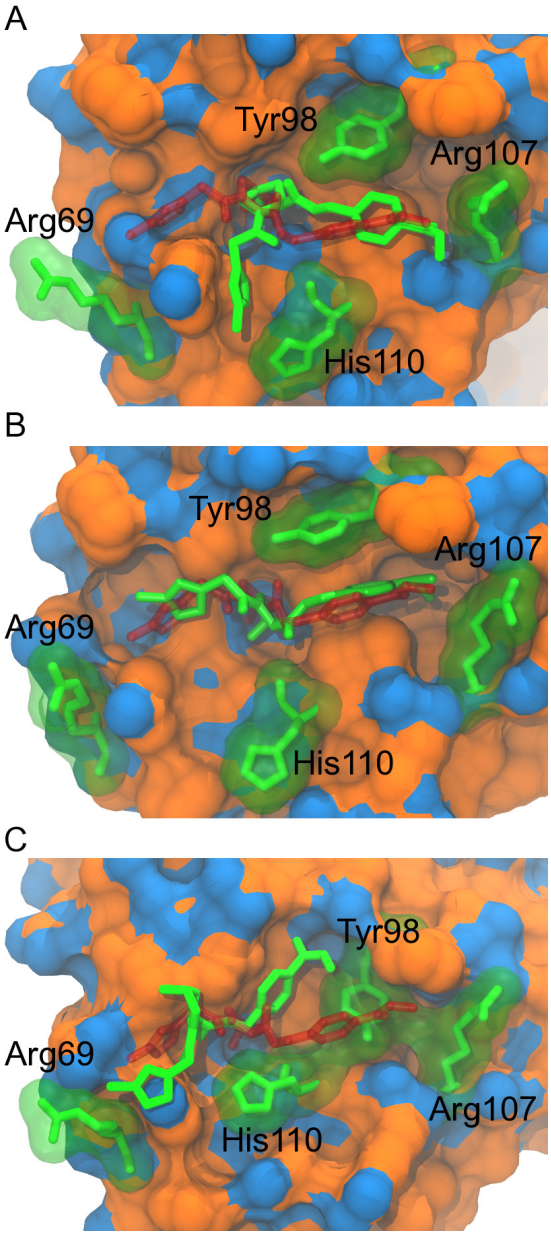

Figure 5.7: Ligand docking in JEDI computed VHL conformations. A) Pose of **1** (green sticks) presenting the lowest RMSD with the ligand in its crystallographic position, docked in the computed apo VHL conformation closest to the average conformation of the most populated cluster from an umbrella sampling simulation with $s_0 = 10.0$ and $\kappa = 2000$ kJ.mol^{-1}.nm^{-2}. B) Same as A) but a more appropriate receptor conformation to bind the ligand has been chosen from the most populated cluster C) Same as A) but docked in the computed apo VHL conformation closest to the average conformation of the most populated cluster from a classical MD simulation. Results obtained using Vina. The crystallographic pose is in red sticks. All other symbols and representations are as in Figure 5.3.

with the crystallographic position of the ligand (RMSD of 3.6 Å), VINA binding energy of -5.6 kJ.mol^{-1}). This is however not the top-scored pose which had a VINA binding energy of -6.2 kJ.mol^{-1}. Qualitatively the discrepancy with the crystallographic binding mode is mostly due to a shift of the isoxazole ring of **1** that is involved instead in stacking interactions with Tyr112. Closer inspection of the computed complex indicates that this binding mode is preferred because the computed 'left-hand side' VHL pocket that would normally host the isoxazole ring is too shallow. However, fluctuations in pocket depth are apparent in snapshots that are present in the same cluster, and it is possible to manually select a snapshot with a left-hand-side pocket that more closely resembles the crystallographic structure. Repeating docking calculations on this conformation (Figure 5.7B) yields indeed a well scored pose (VINA binding energy -6.4 kJ.mol^{-1}) that reproduces fairly well the crystallographic position of the ligand (RMSD of 2.1 Å) though this is again not the top-scoring pose which had a VINA binding energy of -7 kJ.mol^{-1}. As a control, the same docking protocol was also applied to the computed apo VHL conformation closest to the average conformation of the most populated cluster from an unbiased classical MD simulation (Figure 5.7C). As expected, the lowest-RMSD pose was significantly different from the crystallographic binding mode of 1 (RMSD of 5.4 Å), VINA binding energy -6.1 kJ.mol^{-1}). The docking calculations were repeated allowing side-chain flexibility of Tyr98, Ile109 but no improvements were observed. This is likely because significant conformational changes involving both side-chain and backbone atoms rearrangements are necessary to form the ligand binding site from the apo protein conformations sampled from the unbiased MD simulation. Conversely, little improvements was seen in the RMSD of the ligands docked into the JEDI computed conformations with the aid of a flexible side-chain docking protocol, presumably because the binding site is already largely formed.

5.3.2 hPNMT

As described in the previous chapter, the human phenylethanolamine N-methyltransferase (hPNMT) is an enzyme involved in the synthesis of epinephrine from norepinephrine using the cofactor S-adenosyl- L-methionine to methylate the amine of noradrenaline. Two different hPNMT inhibitors, **3** and **4** (Table 5.1), have been identified with a K_i of 0.28 μM and 0.063 μM respectively obtained by radiochemical assay.[16] It has been shown that these two ligands bind two different conformations of the hPNMT binding site (Figure 5.8). Both compounds perform hydrophobic interactions but only the larger ligand (**3**) shows the ability to make hydrogen bonds with the side chain of Lys57 creating a new subpocket which is hidden in the complex with the smaller inhibitor (**3**). The JEDI predictions were able to capture a favorable increase in druggability due to a rise of the hydrophobicity and the enlargement of the cavity due the motion of Lys57 at the edge of the binding site leading to a better druggability (Table 5.1, second row).

5.3.2.1 Molecular Dynamics Simulations

hPNMT was selected as a case study to explore the JEDI druggability predictions in the context of buried cavities. Apo and holo classical MD simulations were performed following the protocol described in Materials & Methods. Results of classical MD trajectories are presented in Figure 5.9. As expected, the simulation achieved in absence of ligand shows much larger $JEDI_{score}$ fluctuations than the

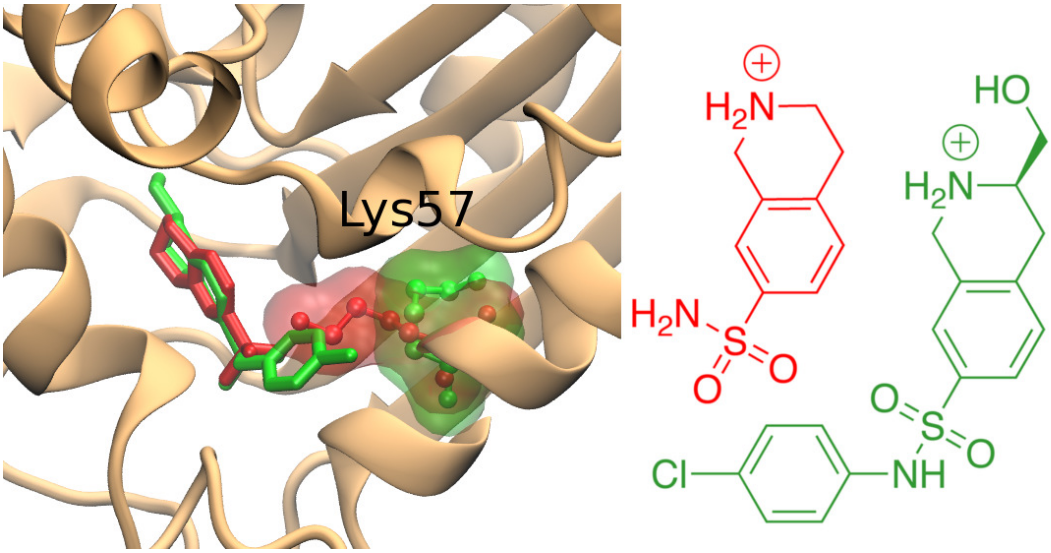

Figure 5.8: Illustration of the hPNMT binding site. The conformational changes inducing an increase of druggability (according to the literature) are highlighted in green (**4**) while the less druggable conformation is represented in red (**3**). The part of the binding site that is not involved in the druggability variation is colored brown. Ligand corresponding to each binding site is represented beside the figure. Pictures were prepared using the software VMD.

simulation of the complex reflecting a destabilization of the holo binding site conformation. Surprisingly, the average $JEDI_{score}$ of the apo simulation (6.67±0.97) is slightly higher than the $JEDI_{score}$ of the holo simulation (6.13±0.71). In addition, those scores are significantly different than the predictions obtained previously using a short MD simulation with position restraints (Table 5.1, first row). This observation is mainly due to an active volume more (V_a) important in the apo simulation (0.15±0.041) than in the holo simulation (0.11±0.038). In contrast with VHL where the $JEDI_{score}$ variations are caused by the motion of few amino acids, backbone motions were observed and Lys57 was not found to be involved in druggability changes. Clustering analysis with a RMSD cutoff of 1 Å reveals the presence of one major binding site conformation (47% of the trajectory) in the holo simulation, that is depicted in Figure 5.9C (right panel).

By contrast, dozens of clusters were observed during the simulation in absence of **3**. The most populated of them (18% of the trajectory) is represented in Figure 5.9C (left panel). The loop between the two *alpha*-helices highlighted in green in Figure 5.9C (top part) was found to be much more flexible in the apo simulation creating frequently enlarged cavities. This backbone flexibility is also described in the less and more druggable conformations observed in the apo simulation (Figure 5.9C, bottom part). Indeed, even if the protein secondary structure is stable along the trajectory, the two α-helices of the N-terminal part of the protein are occasionally occupying the binding site inducing a drop in the $JEDI_{score}$. An increase in JEDI druggability predictions occurred in presence of the ligand after 38 ns (7.3±0.68) and remains stable for the rest of the simulation. This change is correlated to the rearrangement of Tyr85 and Tyr40 in the binding site. First, Tyr85 moves slightly away from the ligand increasing temporally the active volume descriptor (35 ns) and allowing to Tyr40 to be more involved in the binding pocket.

5.3.2.2 Umbrella Sampling Simulations

Umbrella sampling simulations were also performed for this system. A first set of apo simulations were performed encouraging the hPNMT binding site to adopt conformations with a $JEDI_score$ of 8.5 (according to Table 5.1, second row). using different κ values (eq. 5.43). Results are presented in Figure 5.10A. As expected fluctuations from the target value decrease with increased κ values. Trajectories obtained with a κ value of 2000 kJ.mol^{-1}.nm^{-2} were selected to perform clustering analysis. One main cluster containing 51% of the protein conformations was found for the apo simulation. An illustration of the snapshot

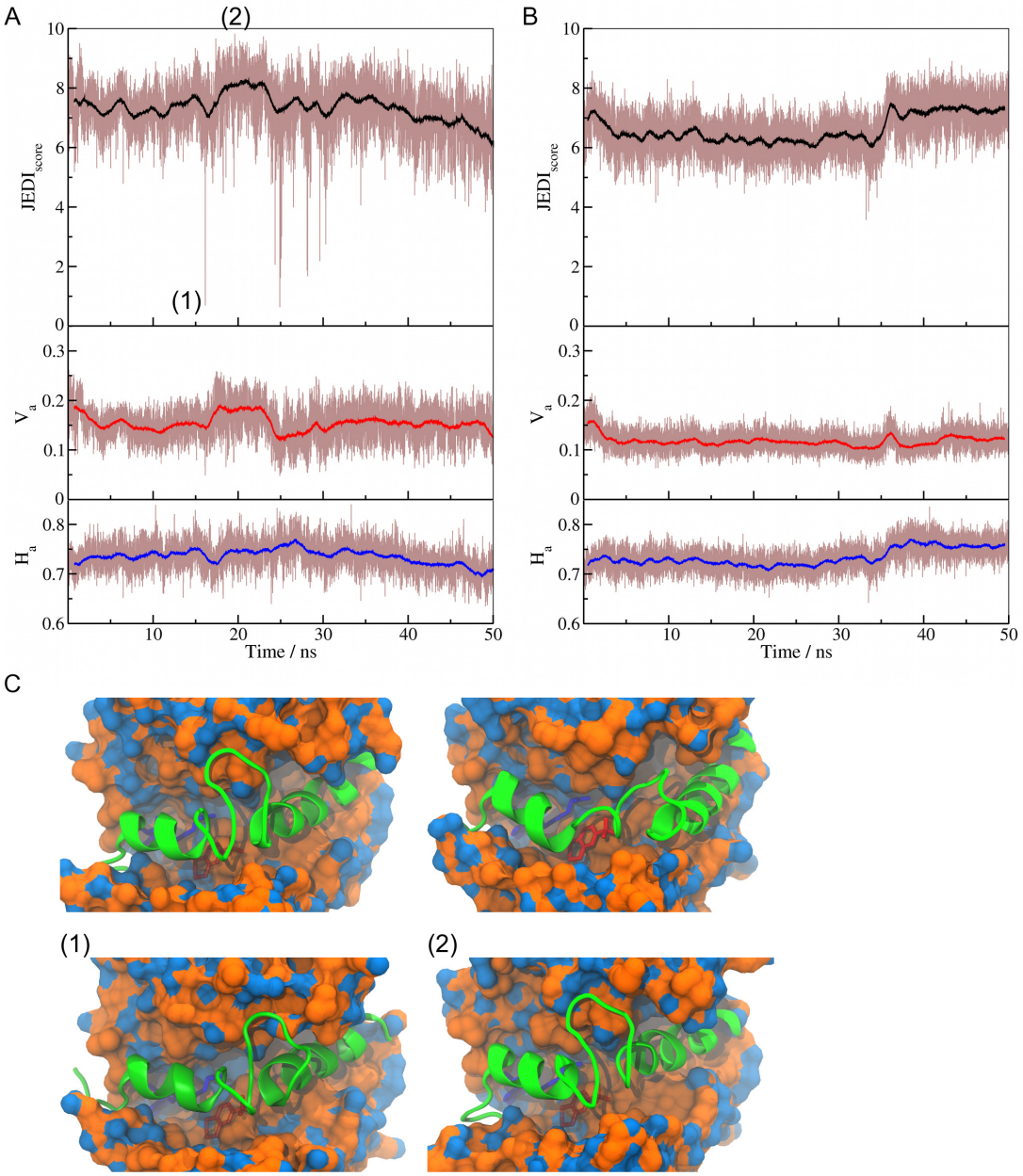

Figure 5.9: Results of the classical MD simulations for hPNMT. The running averages of the changes in $JEDI_{score}$, active volume descriptor and hydrophobicity descriptor are respectively represented in black, red and blue for apo (A) and holo simulation (B). C) The most representative conformations of the apo (top left) and holo trajectory (top right) are shown in surface. The conformations indicated by numbers in figure B are depicted in the bottom part of the figure C. Protein surface were colored according to polar (blue) and apolar atoms (orange). The amino acids responsible for $JEDI_{score}$ variations discussed in the text are highlighted in green. **3** is represented in red. Concerning the apo simulations, the ligand (transparent) was positioned using a structural alignment with the holo crystallographic structure. Pictures were prepared using the software VMD.

closest to the average structure is presented in Figure 5.10A. In contrast with the classical MD simulation, the *alpha*-helix in the N-terminal part (left-hand side) is much less buried leading to the formation of a pocket more suitable to bind a drug-like compound. The hPNMT/**3** umbrella sampling simulations forcing the system to adopt more druggable conformations revealed a partial destabilization of the α-helices increasing the active volume of the binding site. The most representative conformation is depicted in Figure 5.10B.

5.3.2.3 Docking

Several docking experiments were performed to evaluate the conformational ensembles computed from the umbrella sampling simulations. Results are presented in Figure 5.11. First, the computed apo hPNMT conformation closest to the average conformation of the most populated cluster from an umbrella sampling simulation with $s_0 = 8.5$ and $\kappa = 2000$ kJ.mol^{-1}.nm^{-2} (Figure 5.11A) was used to dock **3**. A pose similar to that observed in the crystallographic (RMSD of 1.7 Å) structure was identified in the top-scored poses. As a control, the same docking protocol was also applied to the computed apo hPNMT conformation closest to the average conformation of the most populated cluster from the classical MD simulation (Figure 5.11B). By contrast with the previous study achieved on VHL, the most representative conformation of the non biased trajectory is also able to bind the ligand as observed in the crystallographic structure.

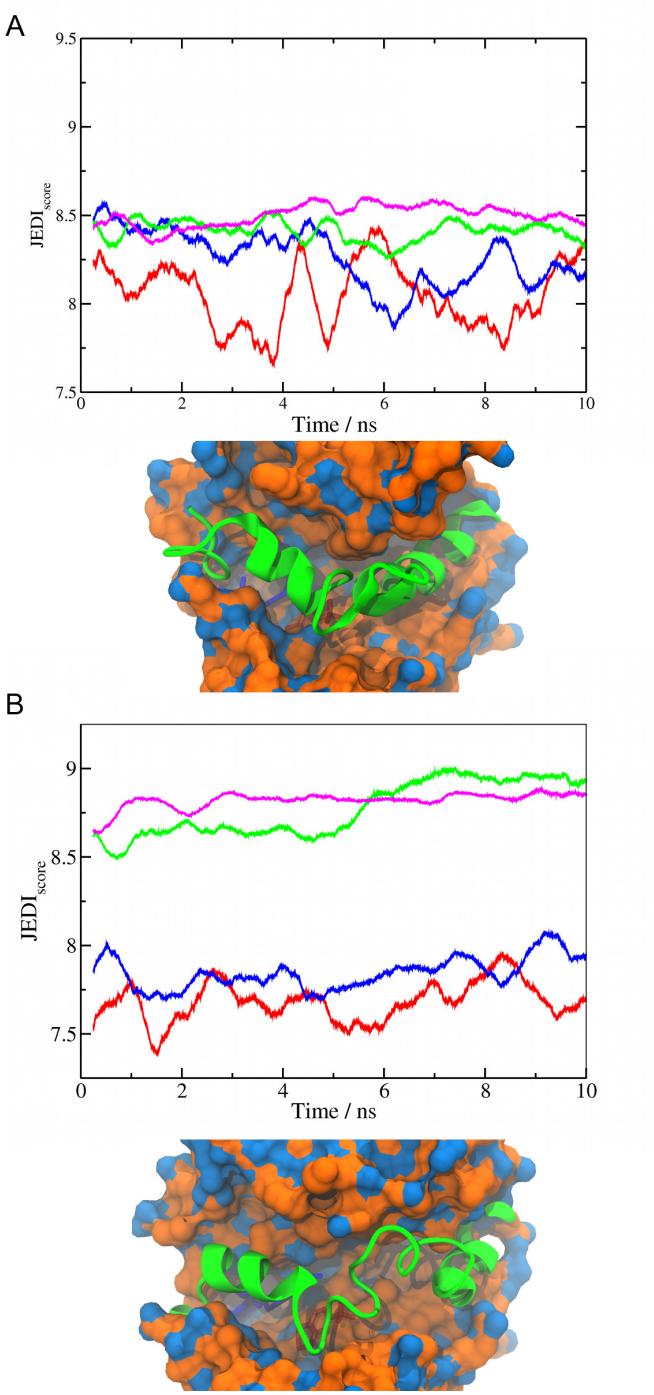

Figure 5.10: Results of the umbrella sampling simulations for VHL in absence (A) and presence (B) of the ligand. For reasons of clarity, only the running averages are shown for four different spring constants (red: $\kappa = 500$ kJ.mol^{-1}.nm^{-2}, blue: $\kappa = 1000$ kJ.mol^{-1}.nm^{-2}, green: $\kappa = 2000$ kJ.mol^{-1}.nm^{-2}, magenta: $\kappa = 5000$ kJ.mol^{-1}.nm^{-2}). An illustration of the most populated cluster of the simulation performed using a force constant of 2000 kJ.mol^{-1}.nm^{-2} is depicted beside each graph. Pictures were prepared using the software VMD.

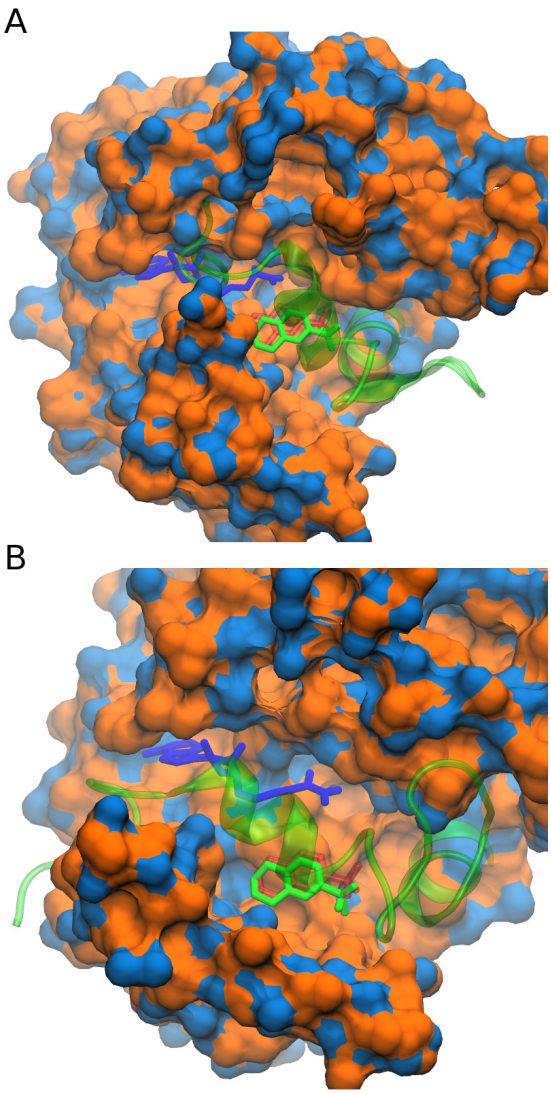

Figure 5.11: Ligand docking in JEDI computed hPNMT conformations. A) Pose of **3** (green sticks) presenting the lowest RMSD with the ligand in its crystallographic position, docked in the computed apo hPNMT conformation closest to the average conformation of the most populated cluster from an umbrella sampling simulation with $s_0 = 8.5$ and $\kappa = 2000$ kJ.mol^{-1}.nm^{-2}. B) Same as A) but docked in the computed apo hPNMT conformation closest to the average conformation of the most populated cluster from a classical MD simulation. Results obtained using Vina. The crystallographic pose is in red sticks. All other symbols and representations are as in Figure 5.3.

5.4 Conclusion

In this chapter, the ability of JEDI to detect cryptic binding sites during classical and biased MD simulations has been investigated. The main novelty of the approach lies in its potential to bias MD simulations with a JEDI force that will encourage a protein region to adopt conformations that match desired druggability scores. The results obtained through several umbrella sampling simulations of VHL indicate that JEDI enables the rapid sampling of 'holo-like' protein conformations that are rarely seen in unbiased apo MD simulations. For structure-based drug design purposes this would be useful to identify tractable conformations in targets that may be otherwise considered undruggable from crystallographic analysis. JEDI also enables biased simulations of protein-ligand complexes. For structure-based drug design purposes, this would be useful to identify enlarged cavities that could accommodate a larger analog of an existing ligand. The results obtained for hPNMT are more contrasted and they highlight limitations of the current implementation of the JEDI approach in the case of buried cavities. The conformational sampling may have been biased by using the GBSA solvent model promoting the formation of compact structures and alpha helix.[17]

5.4 Bibliography

[1] Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provasi, Paolo Raiteri, Davide Donadio, Fabrizio Marinelli, Fabio Pietrucci, Ricardo A Broglia, and Michele Parrinello. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications*, 180(10):1961–1972, October 2009.

[2] Peter P Schmidtke and Xavier X Barril. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *Journal of Medicinal Chemistry*, 53(15):5858–5867, August 2010.

[3] Maestro, version 9.7, Schrödinger, LLC, New York, NY, 2014.

[4] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, March 2008.

[5] Di Qiu, Peter S Shenkin, Frank P Hollinger, and W Clark Still. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *The Journal of Physical Chemistry A*, 101(16):3005–3014, April 1997.

[6] Alexey Onufriev, Donald Bashford, and David A Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*, 55(2):383–394, May 2004.

[7] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, January 2007.

[8] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L Klepeis, Ron O Dror, and David E Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, 78(8):1950–1958, June 2010.

[9] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, July 2004.

[10] Alan W Sousa da Silva and Wim F Vranken. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC research notes*, 5:367, July 2012.

[11] Junmei Wang, Wei Wang, Peter A Kollman, and David A Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics & Modelling*, 25(2):247–260, October 2006.

[12] G M Torrie and J P Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, February 1977.

[13] Oleg Trott and Arthur J Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, January 2010.

[14] Daniel Seeliger and Bert L de Groot. Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *Journal of Computer-Aided Molecular Design*, 24(5):417–422, May 2010.

[15] Carles Galdeano, Morgan S Gadd, Pedro Soares, Salvatore Scaffidi, Inge Van Molle, Ipek Birced, Sarah Hewitt, David M Dias, and Alessio Ciulli. Structure-Guided Design and Optimization of Small Molecules Targeting the Protein-Protein Interaction between the von Hippel-Lindau (VHL) E3 Ubiquitin Ligase and the Hypoxia Inducible Factor (HIF) Alpha Subunit with in Vitro Nanomolar Affinities. *Journal of Medicinal Chemistry*, 57(20):8657–8663, October 2014.

[16] Gary L Grunewald, Mitchell R Seim, Rachel C Regier, and Kevin R Criscione. Exploring the active site of phenylethanolamine N-methyltransferase with 1,2,3,4-tetrahydrobenz[h]isoquinoline inhibitors. *Bioorganic & Medicinal Chemistry*, 15(3):1298–1310, February 2007.

[17] Ruhong Zhou. Free energy landscape of protein folding in water: Explicit vs. implicit solvent. *Proteins*, 53(2):148–161, September 2003.

6

Conclusion

This chapter gives an overview of the concepts presented throughout the thesis and discusses future development of the JEDI approach

The work presented throughout this thesis aims to propose a new methodology to strengthen the reliability of computer-aided structure-based drug design. Computational approaches have become an increasingly important part of the drug discovery process.[1] Besides the cost reduction in terms of human resources, time and money, the growing number of three-dimensional (3D) protein structures available in platforms such as the Protein Data Bank (PDB) is an additional motivation to the development of new bioinformatics and chemoinformatics tools.

Protein flexibility is essential in many aspect of cellular biochemistry. In solution, a protein can not be considered as a static entity and may adopt structurally different conformations of similar low energies. For instance, intrinsically disordered proteins (IDPs) can adopt a broad range of conformations, ranging from collapsed to fully extended. The considerable flexibility of IDPs facilitates interactions with a large number of proteins and explains why IDPs often play a crucial role in important cellular processes such as signaling or transcription.[2, 3] IDPs are attractive therapeutic targets as they are often implicated in a broad range of diseases, such as cancers, cardiovascular disease or neurodegenerative diseases. However, the considerable flexibility of IDPs presents a challenge for drug discovery approaches.[4] The interactions of small molecules with IDPs challenge our understanding of molecular recognition and it is important to clarify the mechanisms of IDP-small molecule interaction before such proteins can be more routinely targeted. In this thesis, interactions regulating the formation of IDP-small molecule complexes have been reviewed through three well-studied systems: the oncoprotein c-Myc, Aβ (amyloid β-peptide) and α-synuclein.[5] They have highlighted the difficulty to develop pharmaceutical compounds able to bind IDPs with high affinity and selectivity.

The oncoprotein c-Myc was selected to perform further analysis in order to provide a better understanding of the interactions characterizing IDP-small molecule complexes. Several classical molecular dynamics (MD) and metadynamics simulations were performed to study the conformational sampling of a c-Myc truncated peptide (c-Myc$_{402\text{-}412}$). Results obtained from simulations performed in absence and presence of a known inhibitor suggest mainly weak interactions between the ligand and the peptide. Moreover, it has been found that c-Myc$_{402\text{-}412}$ remains partially disordered upon the binding of the small molecule. Therefore, many protein conformations were observed making identification of hidden pockets very difficult. According to the literature a few well defined pockets were expected. For this reason, simulation techniques were used to compute the conformational ensemble of c-Myc$_{402\text{-}412}$ in order to characterize these pockets. However, many protein conformations were observed making identification of hidden pockets very difficult. Therefore, those results highlight the difficulty to propose guidelines to help the optimization of ligands binding IDPs into more potent inhibitors.

In the last decades, a lack of druggability was found to be one the major causes of failure in the drug discovery process.[6, 7] This thesis introduced a new computational approach aiming to identify hidden pockets at protein surfaces and characterize their druggability. JEDI was designed to capture binding site druggability fluctuations during a MD simulation. The main novelty of this methodology is the possibility to use the JEDI scoring function as a collective variable to bias MD simulations. Indeed, druggability predictions are computed using a potential that is fast, continuous and differentiable. In addition to

the ability to distinguish druggable from nondruggable binding sites, other characteristics such as the computational cost and the sensitivity were investigated. The methodology was found to be as accurate as alternative approaches to discriminate nondruggable from druggable binding sites. In addition, JEDI is fast enough to perform classical and biased MD simulations within reasonable time.

Simulations performed on a solvent exposed binding site (VHL) have highlighted interesting perspectives for structure-based drug design purposes such as the identification of new druggable binding site conformations or ligand optimization. However, the current implementation of JEDI only allows simulations in implicit solvent that may cause sampling issues as it has been observed with the second case study (hPNMT). Further work will focus on replacing the GBSA implicit solvent model with explicit solvent models, and this is expected to improve the accuracy of the computed conformations.[8] Clustering of the biased simulations in VHL has identified in many instances several structurally distinct conformations that match a given target druggability value. That druggability is a degenerate collective variable is not unexpected. An exciting direction for this work is to couple the JEDI calculations with other collective variables to resolve the distinct hidden conformational states. This will facilitate the evaluation of the free energy of these hidden conformational states with respect to the native state conformation. This parameter is likely to be important for practical applications. Presumably the feasibility of targeting productively with a ligand a putative hidden binding site hinges on an acceptable stability relative to the native state.[9] Another interesting perspective could be to make JEDI grid calculations on entire protein structures to identify binding sites and assess their druggability. Such 'blind detections' may require clustering approaches to characterize grid point connectivity and identify distinct binding pockets.

6.0 Bibliography

[1] Izeth M Kapetanovic. Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. *Chemico-Biological Interactions*, 171(2):165–176, January 2008.

[2] Lilia M Iakoucheva, Celeste J Brown, J David Lawson, Zoran Obradovic, and A Keith Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of Molecular Biology*, 323(3):573–584, October 2002.

[3] Andrew Campen, Ryan M Williams, Celeste J Brown, Jingwei Meng, Vladimir N Uversky, and A Keith Dunker. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein and Peptide Letters*, 15(9):956–963, 2008.

[4] Vladimir N Uversky, Christopher J Oldfield, and A Keith Dunker. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annual Review Biophysics*, 37:215–246, 2008.

[5] Rémi Cuchillo and Julien Michel. Mechanisms of small-molecule binding to intrinsically disordered proteins. *Biochemical Society transactions*, 40(5):1004–1008, October 2012.

[6] David Brown and Giulio Superti-Furga. Rediscovering the sweet spot in drug discovery. *Drug Discovery Today*, 8(23):1067–1077, December 2003.

[7] Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, 3(8):711–715, August 2004.

[8] Ruhong Zhou. Free energy landscape of protein folding in water: Explicit vs. implicit solvent. *Proteins*, 53(2):148–161, September 2003.

[9] Julien Michel. Current and emerging opportunities for molecular simulations in structure-based drug design. *Physical Chemistry Chemical Physics*, 16(10):4465–4477, March 2014.

A

Appendix

List of Publications:

Publication 1.
Julien Michel & Rémi Cuchillo. The Impact of Small Molecule Binding on the Energy Landscape of the Intrinsically Disordered Protein C-Myc, *PLOSone*, 7(7): e41070, 2012.

Publication 2.
Rémi Cuchillo & Julien Michel. Mechanisms of Small-Molecule Binding to Intrinsically Disordered Proteins, *Biochemical Society Transactions*, 40, 1004-1008, 2012.

Publication 3.
Rémi Cuchillo, Kevin Pinto-Gil & Julien Michel. A Collective Variable for the Rapid Exploration of Protein Druggability, *Journal of Chemical Theory and Computation*, 2015 (in press).

List of Conferences:

CCP5 summer school, Cardiff, 2012.
Poster: The impact of small molecule binding on the energy landscape of the intrinsically disordered protein c-Myc.

CCPBioSim, Nottingham, 2013.
Poster: A Collective Variable for Rapid Exploration of Protein Druggability.

Young Modellers' Forum, London, 2014.
Oral presentation: Protein druggability: the JEDI approach.

CCPBioSim, Edinburgh, 2014.
Poster: Protein druggability: the JEDI approach.