Declaration of Authorship

I, Michail PAPADOURAKIS, declare that this thesis titled, "Molecular Dynamics Based Methods for the Computation of Standard Binding Free Energies and Binding Selectivity of Inhibitors of Proteins of Pharmaceutical Interest" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Papadourakis Michail

Date: 29th of October 2020

Abstract

The field of Computer Aided Drug Design (CADD) has experienced substantial developments over the last few decades thanks to a rapid growth in computing power. In particular, Molecular Dynamics (MD) simulations and associated techniques have earned increased attention within the pharmaceutical sector thanks to their rising accuracy and diminishing cost. However, there are still limitations in the usage of these methods, due to the difficulty of sampling the rugged energy landscapes of protein-ligand complexes. The main theme of this work is to address the sampling problem of MD methods for predicting the binding free energies of different biomolecular complexes.

This work starts using MD simulations as a sampling technique for a relative free energy calculation protocol using the Sire Open Molecular Dynamics (SOMD) software. This protocol was then integrated in a ligand design workflow to optimize the binding selectivity of cyclophilin (Cyps) inhibitors. Cyps are proteins known to play a vital role in various diseases, such as cancer, Alzheimer and viral infections. Most Cyp inhibitors to date, however, are cyclic peptides that have potency in the nanomolar range but produce severe side effects, are complex to synthesize and display complex pharmacokinetic profiles. Thus, there is a need for new selective small molecules targeting specific Cyps isoforms, in order to gain new insights for the inhibition of these therapeutically vital proteins. The computational workflow was able to suggest auspicious designs that they will be synthesized and characterized using biophysical techniques from Alison Hulme's lab.

Following, MD simulation methods were employed for the more challenging task of predicting the absolute free energies of binding of protein-ligand complexes. For this purpose, an Alchemical Free Energy (AFE) protocol was generated and its efficiency was evaluated in the Statistical Assessment of Modelling of Proteins and Ligands (SAMPL6) challenge. SAMPL challenges involve a series of blinded predictions of standard binding free energies for toy host-guest molecules. The results obtained from our protocol were ranked among the top submissions in terms of accuracy and correlation with experimental data.

Encouraged by these results, we wanted to compare the efficiency of the AFE protocol versus a Markov State Modelling (MSM) protocol for the calculation of the standard binding free energy of a ligand to the intrinsically disordered protein c-Myc. The oncoprotein c-Myc is overexpressed in over 70% of human cancers and its inhibition has been considered the holy grail in cancer therapy. Due to its structural elasticity it is difficult to perform structure-based drug design methods for the discovery of novel compounds. The results showed that MSM can describe accurately the binding process of the ligand to the oncoprotein c-Myc, but the binding free energies were similar with the ones of the AFE protocol.

Finally, an adaptive sampling protocol was established for the computation of the standard binding free energy and binding selectivity of lead-like ligands for the flexible protein MDM2. MDM2 is a vital protein that acts as an inhibitory mechanism of the transcription factor p53. p53 plays an important role in the regulation of cellular processes and suppression of tumor development. For this reason, it is important to develop methods for the discovery of novel ligands that could inhibit the MDM2-p53 interaction through binding to the MDM2 protein. The results of the adaptive sampling study were encouraging as the protocol was able to predict binding selectivity trends for the MDM2-ligand complexes approximately six times faster than the original AFE protocol.

Lay summary

The work in this thesis describes the development and use of computational methods to simulate protein-ligand complexes. The goal is to validate the utility of the simulations for drug discovery research.

The purpose of drug design is to identify a compound (ligand) that binds to a protein or enzyme (biomolecule) involved in a disease to inhibit its biological function. An important goal in computer-aided drug design is to predict how efficiently and selectively a ligand binds to a target protein, as most drugs must be potent and selective binders to effectively inhibit the function of a protein target, whilst avoiding undesirable side-effects due to binding to other proteins. Despite decades of efforts, making such predictions routine remains challenging.

One method that has proven popular to improve the drug design is Molecular Dynamics (MD) simulations. This technique uses Newton's equations of motion to simulate the movement of atoms and molecules. There are many successful examples of usage of this method to design ligands for proteins. However, the techniques suffers still from limitations. A major problem is how well this method samples all the different conformations that a protein-ligand complex can adopt. This work explores this problem by using MD simulations in different protocols and workflows to compute the strength of the interactions of drug-like molecules targeting different proteins of interest for the pharmaceutical industry.

Acknowledgements

First of all, I would like to express my gratitude to my supervisor Dr. Julien Michel for his guidance, support, knowledge and encouragement throughout my PhD research. Thanks also to my second supervisor, Prof. Philip J. Camp, who helped me to tackle Vivas with his hard questions and his support. I want also to thank Dr. Zoe Cournia for the opportunity that she gave me to work on her lab during my second year of studies. In addition, I want to thank HPC-Europa3 scholarship for covering my travel bursaries and providing me a stipend during my stay in Greece.

Many thanks to all the current and past members of the 234 office in the school of Chemistry for their suggestions and discussion in research and the warm environment in our office. Special thanks to Dr. Stefano Bosisio that trained me during my first year of the PhD, Cyclophilins' team (Alessio, Arun, Jordi, Harris, Charis, Maria) for their help and discussions for this project and Toni for her support at the c-Myc project. Also, I am grateful to Jordi who helped me and trained me throughout the 3 years of my PhD.

In addition, I would like to thank Lavrentis and Harris for their support and their friendship during my stay in Edinburgh. I would like to thank my girlfriend Maria that helped me and encouraged me in every decision. Thanks for our laughs and your love.

Finally, many thanks to my mom, dad and sister for understanding my choices and for your support. Without you I would never had the chance to start my PhD education journey.

Contents

List of Figures

List of Tables

List of Abbreviations

ADMETox	Absorption Distribution Metabolism Excretion and Toxicology
AFE	Alchemical Free Energy
ALS	Amyotrophic Lateral Sclerosis
Amber	Assisted Model Building with Energy Refinement
aMD	accelerated Molecular Dynamics
AP	Accute Pancreatosis
ATP	Adenosine 5′-TriPhosphate
BAR	Bennet Acceptance Ratio
BEMD	Bias-Exchange variant of the MetaDynamics method
bHLHZip	basic-Helix-Loop-Helix-Leucine Zipper
Bzd	1,4-Benzodiazepine-2,5-dione
CAAD	Computer Aided Drug Design
CD	Circular Dichroism
CK	Chapman-Kolmogorov
CNS	Central Neural System
CsA	Cyclosporin A
CsC	Cyclosporin C
CSD	Cambridge Structural Database
Cyps	Cyclophilins
DFT	Density-Functional Theory
E-boxes	Enhancer Boxes
ER	Endoplasmic Reticulum
$\mathbf{F}_{NVT}$	Helmholtz Free Energy
FDA	Food and Drug Administration
FDTI	Finite Difference Thermodynamic Integration
FEP	Free Energy Perturbation
FRET	Fluorescence Resonance Energy Transfer

$\mathbf{G}_{NVT}$	Gibbs Free Energy
GAFF	General AMBER force field
Gromacs	GROningen MAchine for Chemical Simulations
h	Plank's Constant
HCV	Hepatitis C Virus
HIV	Human Immunodeficiency Virus
$\mathbf{H}_i(\mathbf{q}^N, \mathbf{p}^N)$	Hamiltonian Function
HTS	High Throughput Screening
$\mathbf{IC}_{50}$	Half Maximal Inhibitory Concentration
ICAM-1	Intercellular Adhesion Molecule-1
IDPs	Intrinsically Disordered Proteins
ITC	Isothermal Titration Calorimetry
ITS	Implied TimeScales
$\mathbf{K}_a$	Equilibrium Association Constant
$\mathbf{K}_d$	Equilibrium Dissociation Constant
$\mathbf{K}(\mathbf{p}^N)$	Kinetic Energy
LBDD	Ligand Based Drug Design
MAPK	Mitogen Activated Protein Kinase
MBAR	Multistate Bennet Acceptance Ratio
MC	Monte Carlo
MD	Molecular Dynamics
MDM2	Murine Double Minute-2
MFTP	Mean First Passage Time
MM/PBSA	Molecular Mechanics/Poisson–Boltzmann Surface Area
MoRE	Molecular Recognition Elements
mPTP	Mitochondrial Permeability Transition Pore
MSM	Markov State Model
NMR	Nuclear Magnetic Resonance
NAFLD	Non-Alcoholic Fatty Liver Disease
NAMD	NAnoscale Molecular Dynamics
NASH	Non-Alcoholic SteatoHepatitis
NPT	Isothermic Isobaric Ensemble

NS5A	NonStructural Protein 5A
NVT	Canonical Ensemble
PBC	Periodic Boundary Conditions
PCCA+	Perron Cluster Cluster Analysis
PD	PharmacoDynamic
PK	PharmacoKinetic
PME	Particle-Mesh Ewald
PPIase	Peptidyl-Prolyl cis-trans Isomerase
$\mathbf{Q}_{id}$	Momentum Integral
$\mathbf{Q}_{NVT}$	Canonical Partition Function
RA	Reumatoid Arthritis
RMSD	Root-Mean Square Deviation
R&D	Research and Development
ROS	Reactive Oxygen Species
SAMPL6	Statistical Assessment of Modelling of Proteins and Ligands
SBDD	Structure Based Drug Design
SFSS	Small for Size Liver Syndrome
SOMD	Sire/OpenMM Molecular Dynamics
SPR	Surface Plasma Resonance
SREBP-1C	Sterol Regulatory Elementary Binding Protein-1C
TAD	Transactivation Domain
TI	Thermodynamic Integration
TICA	Time-lagged Independent Component Analysis
TNF	Tumor Necrosis Factor
$\mathbf{U(q}^{N}\mathbf{)}$	Potential Energy
US	Umbrella Sampling
vdW	van der Waals
vFEP	variational Free Energy Profile
Vpr	Viral protein R
VS	Virtual Screening
XED	eXtended Electron Distribution
$\mathbf{Z}_{NVT}$	Configurational Integral
Δ_{NVT}	Partition Function

Δ **G**	Binding Free Energy
ΔG°	Standard Binding free energy
ΔG_{rest}	Standard State Correction Term
Λ	Thermal de Broglie Wavelength

List of Publications

Work from this thesis has been published in the following papers:

1. **Papadourakis Michail**, Stefano Bosisio, Julien Michel, "*B*linded predictions of standard binding free energies: lessons learned from the SAMPL6 challenge" *J*ournal of Computer-Aided Molecular Design **32**, 1047-1058 (2018).
2. Andrea Rizzi, Travis Jensen, David R. Slochower, Matteo Aldeghi, Vytautas Gapsys, Dimitris Ntekoumes, Stefano Bosisio, **Papadourakis Michail**, Niel M. Henriksen, Bert L. de Groot, Zoe Cournia, Alex Dickson, Julien Michel, Michael K. Gilson, Michael R. Shirts, David L. Mobley, John D. Chodera, "*T*he SAMPL6 SAMPLing challenge: assessing the reliability and efficiency of binding free energy calculations" *J*ournal of Computer-Aided Molecular Design **34**, 601-633 (2020).

Chapter 1

Introduction

1.1 The Drug Discovery Process

1.1.1 Historical overview of drug discovery

Since the early days of human civilization, there has been a constant need for therapeutic intervention for the treatment of diseases. However, until the 19th century the drug discovery process was mainly based on traditional medicines and natural remedies that were usually discovered by serendipity.[1] For modern drug discovery to evolve, it was important for key scientific fields such as chemistry and pharmacology to advance and mature as sciences.[2] During the 19th century, Avogadro's law, the establishment of the periodic table and the categorisation of chemical compounds as acid and bases opened the way for the rapid development of chemistry. These breakthroughs together with Kekule's benzene theory led to the evolution of dye chemistry. Dyes were already discovered by accident in 1856 when William Henry Perkins synthesised the first synthetic dye, tyrian purple, and established his own factory the following year. As the industrial revolution was taking place during this time, other companies were also founded with the purpose of developing novel synthetic dyes by optimising Perkins' reaction. The rapid growth of this field after benzene theory, made large dye companies such as Bayer and Sandoz realise that organic molecules could be employed as possible drugs and this led to the birth of the pharmaceutical industry.[3, 4]

Apart from their important role on the genesis of the pharmaceutical sector, dyes were also the starting point for modern drug discovery methods. Paul Elrich, a student in the laboratory of the anatomist Wilhelm Waldeyer at the University of Strasbourg, studied the differential affinities of dyes for biological tissues. This lead him to theorize the existence of "chemoreceptors" that affect the interaction of cells with small molecules and produce a biological effect. He further postulated the "magic bullet" theory, where certain chemoreceptors of cancer cells or infectious organisms would be different from analogous structures of the host and these differences could be exploited therapeutically. These findings were fully validated from the identification of *Salvarsan*, the first synthetic arsenic compound able to treat syphilis. This successful paradigm formed the basis of modern chemotherapy in the 20^{th} century. Furthermore, substantial advances in analytical chemistry during the 19^{th} century added new tools to medicine in the 20^{th} century by enabling the purification and characterisation of active ingredients from plants and other extracts, such as the isolation of morphine from opium extract by F. W. Serturner.[5]

All these advances in chemistry, biology and pharmacology laid the foundations for the first reliable biological screening and evaluation pipelines. The most famous example from this procedure was the discovery of penicillin in 1928 by Alexander Fleming.[6] Because of its efficacy and lack of toxicity, penicillin was made available on large scale and helped Allied soldiers in World War II overcome bacterial infections. After the Second World War researchers concentrated on understanding, through experiments on animals, the mechanism of action of possible drug candidates. This led to the development of drugs with similar pharmacokinetic properties to penicillin, such as sulfactams and to the discovery of molecules with different therapeutic effects such as diuretics.[7] In addition, rapid technological development during the 1980s enabled the relatively young field of computational chemistry to bear on drug discovery processes. This field has opened the door for understanding in atomic details protein-ligand interactions[8, 9] and permitted simulations of very large biomolecules to accelerate the drug discovery process.[10] At the same time, the development of more sophisticated *in vitro* experiments, laid the foundations of more accurate and rapid screening of the chemical space.

1.1.2 Modern drug discovery

As a result, nowadays, drug discovery has become an intersection of several scientific disciplines that are constantly evolving such as genomics and proteomics, biology, medicinal and computational chemistry, pharmacology, clinical medicine and biotechnology. However, the development of a new drug remains a complex process that can take up to 15 years of work and Research and Development (R&D) costs in excess of a 1 billion dollars.[11]

The drug discovery process is usually divided into five main steps. The first one is the identification of a biomolecular target using genomics and proteomics and wherever possible the isolation of this target through X-ray crystallography. Once the pharmaceutical target is identified, validated and isolated, thousands of compounds are screened against it using *in vitro* experiments. From these molecules, the identification and optimisation of a handful of lead compounds by medicinal chemists can be assisted using tools provided by the computational chemistry field. These tools can provide insight on the strength of the binding between the lead compounds and the biomolecular target (free energy of binding) and on the physicochemical properties of the compound of interest (drug-like properties). Then, the preclinical stage is initiated in order to understand the Absorption Distribution Metabolism Excretion and Toxicology (ADMETox) profile of the lead compounds through *in vivo* assays before authorising them for human clinical trials. The following clinical development consists of three phases with increased number of volunteers in each phase. At the end of this process, one compound that passes all the tests is approved by drug administration agencies such as Food and Drug Administration (FDA). After its approval, the molecule gains the status of drug and is released in the corresponding market.[12] An overview of this pipeline is illustrated in the following Figure.

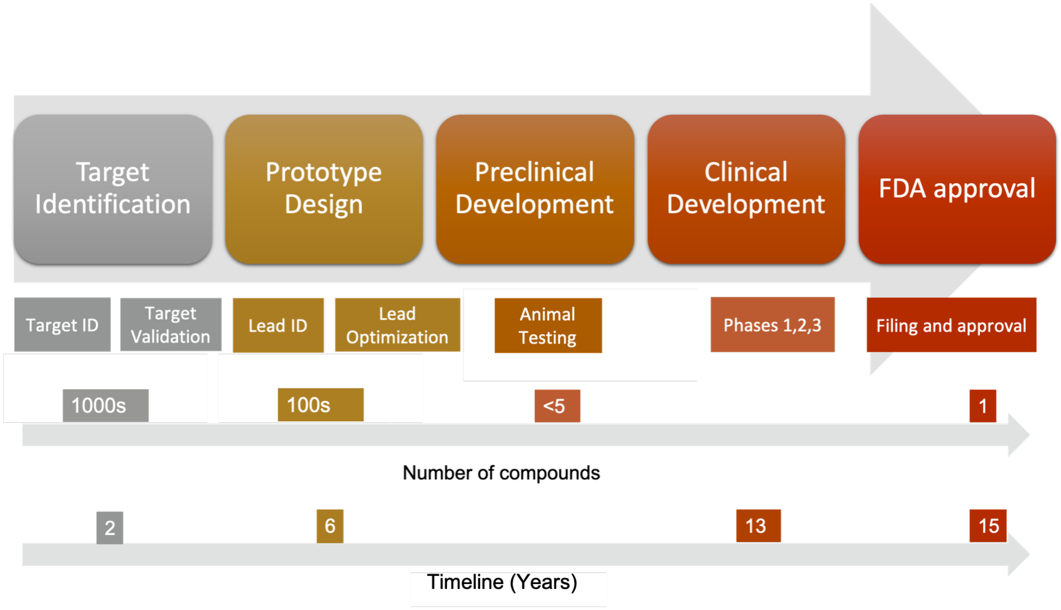


FIGURE 1.1: Traditional drug discovery process. The process starts with the biological research, where a biomolecular target is identified using proteomics and it is isolated using crystallographic techniques. After the identification of the target a library of thousands of drug candidates against this target is screened. Then, the prototype design stage uses tools from computational chemistry, to predict drug-like properties and to define lead compounds. In this stage, the compounds undergo a lead optimization process, which should give a few lead molecules (around five) ready for the preclinical stage. This stage is required to understand the ADME properties of the lead compounds through animal testing, before authorizing these molecules for human trials. The last part of the drug discovery process is devoted to the clinical development, which follows three phases where drugs are tested on gradually larger number of volunteers. The final step is the approval by a drug administration organization, such as the FDA, and the launch on the market. The overall process takes approximately 15 years on average.

1.1.3 Structure Based Drug Design

The present work deals with issues encountered in the prototype design step of the drug discovery process. As mentioned above computational methods are frequently used to help medicinal chemists identify a small number of lead compounds that will then proceed to preclinical and clinical evaluation. The Computer Aided Drug Design (CAAD) tools used for this purpose depend on the information available for the biomolecule target and the drugs used for its inhibition. If only the ligands that inhibit the target are known, then a Ligand Based Drug Design (LBDD) approach is applied, whereas if the structure of the pharmacological target is available then the drug design process can be Structure Based (SBDD).[13, 14]

The SBDD method is an auspicious approach for drug design, where promising drug candidates are chosen based on the structural information derived from the experimental data. It is an iterative process, as it proceeds through multiple cycles before an optimised lead is designed. The first step of this approach is to determine experimentally the 3D structure of the therapeutically important protein using NMR, X-ray crystallography, or cryo-electron microscopy. If the experimental observation of the structure is not possible, then computational methods can be used to model the protein's 3D structure. In particular, homology modelling is one of the most reliable approaches to predict the 3D structure of a therapeutically important protein. Homology modelling software, such as SWISS-MODEL, construct an atomic resolution model of the targeted protein using its amino acid sequence and a known homologous protein with >40% similarity.[15, 16]

Once a 3D model of the target is available, attention turns to the location of binding pockets that are suitable to modulate the function of the target. This can be apparent if the model contains a ligand, or if the goal is to block access to a previously characterised functional site. If potential binding sites are not immediately apparent there are several methods that can spot potential binding sites that can interact favorably with important functional groups on possible drug candidates.

Once the structure and the binding site are identified, the next step is the hit discovery. A hit compounds is a molecule that demonstrates promising therapeutic activity at a given protein target. There are two main paths for the discovery of hit compounds that are classified as computer aided versus experimental. The latter path is performed through high throughput screening (HTS) or fragment-based screening techniques, in which a plethora of compounds are tested for biochemical effects or binding effects. On the other hand, the former path uses Virtual Screening (VS) approaches, where libraries of available small molecules are docked into the binding site and ranked by binding affinity.[17] The most promising hit compounds from these methods are then further tested for binding affinity on the target protein using different biophysical assays such as Surface Plasmon Resonance(SPR) and Isothermal Titration Calorimetry (ITC) or using biochemical assays.

The molecules with low μM binding activity, as measured from the aforementioned assays, undergo an iterative hit-to-lead process. In each cycle, the results from the biological assays are analysed to identify promising interactions between the ligand and the protein. These interactions are used to design and synthesise new compounds. The same cycle is then repeated multiple times, until the binding affinity and the selectivity of a small number of compounds (lead compounds) is optimal. In parallel or after the lead identification pipeline, a similar procedure is taking place. In the lead optimisation stage, the lead compounds are extensively and iteratively optimised in order to improve the ADMETox properties while maintaining their potency.

Over the years computer-aided drug design (CADD) has become established as a powerful methodology for the drug-discovery process.[18] CADD can be used to predict the ADMETox properties of the lead compounds. Another application of CADD is the use of molecular simulations methods to support drug-discovery efforts[19], via, for instance, the investigation of protein folding mechanisms[20, 21] or ligand modulation of millisecond time-scale conformational changes in proteins.[22] A major application of CADD is in potency predictions to decrease time and costs of the aforementioned hit-to-lead and lead optimization stages.[23] Computational methods can help with the description, characterization and quantification of the energetics that govern protein-ligand complex formation.

1.2 Models of Molecular Recognition

Molecular recognition is the process of the specific interaction between biological macromolecules (proteins) and small molecules (ligands) through noncovalent bonding, for instance electrostatic or van der Waals interactions. This process is defined by *specificity*, which is the ability of a protein's binding site to bind specific molecules and by *affinity*, which determines the strength of the binding interaction between the protein and the ligand.[24] There are three conceptual models that can describe this procedure in protein-ligand binding. The first model is the *lock-and-key model* where protein and ligand are rigid, and only a correct shaped ligand (key) can bind to the binding site (key hole) of the protein (lock).[25] However, this approach can not explain the experimental finding that a protein can bind its ligand without their initial shapes being complementary. Thus, a second theory was derived, *induced fit model*, which postulates that the protein's binding site is flexible and can undergo a conformational change when interacting with ligand.[26] This approach is ideal for proteins with minor conformational changes, as it accounts only the flexibility of the ligand-binding site. However, the vast majority of proteins are dynamic in nature and form an ensemble of conformational states. Therefore, a third model is introduced, namely the *conformational selection model*, to theorise that the ligand will bind selectively to the most suitable conformational state and will shift the equilibrium towards this state.[27] Finally, there is also a class of proteins that can have little or no tertiary structure and thus a fourth model is required to describe their binding mechanism with small molecules. This model is called *folding-upon-binding* and describes the folding of these proteins into ordered structures when they interact with small ligands.[28] An overview of the four underlying mechanisms is depicted in Figure 1.2.

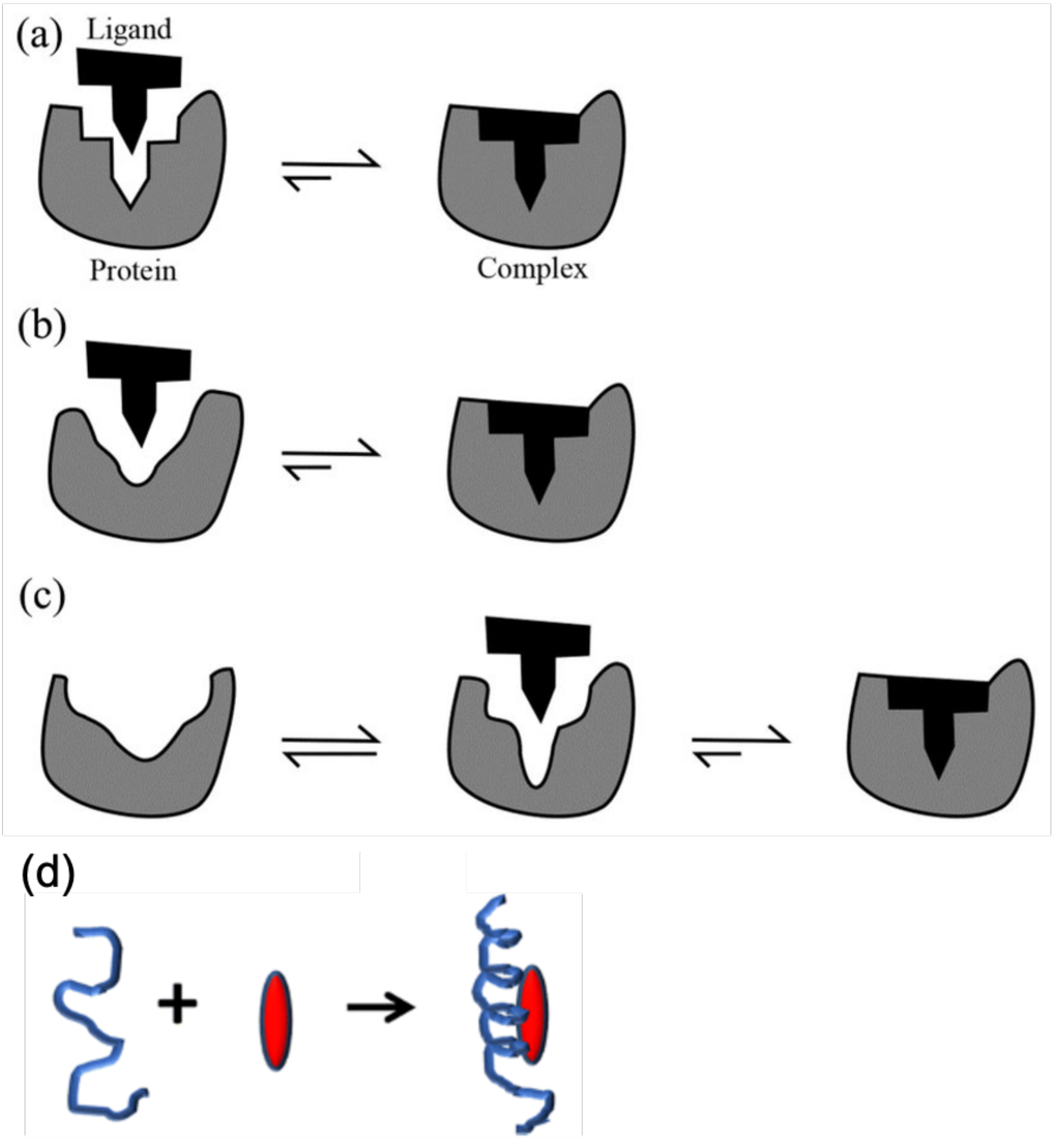


FIGURE 1.2: The four protein-ligand binding mechanisms (a) Lock-and-key; (b) Induced fit; (c) Conformational selection; and (d) Folding-upon-binding. Adopted from paper Du *et al* and Castano *et al*. [24], [29]

Despite the mechanisms that interpret the protein-ligand binding, it is also important to elucidate the physicochemical mechanisms underlying the protein–ligand interaction. The reversible binding of a ligand L to a protein P can be written as in Equation 1.1:

$$P + L \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} PL\,, \tag{1.1}$$

where PL is the protein-ligand complex, k_{on} and k_{off} are the rate constants of binding and unbinding with units s^{-1} M^{-1} and s^{-1} respectively. The above binding reaction should be balanced by the reverse unbinding reaction and this is written as:

$$k_{on}[P][L] = k_{off}[PL]\,, \tag{1.2}$$

where *[P]*, *[L]* and *[PL]* are the equilibrium concentrations of the protein, ligand and the protein-ligand complex respectively. The binding constant K_b (units M^{-1}) is linked with the dissociation constant K_D (units M) through the following relationship:

$$K_b = \frac{k_{on}}{k_{off}} = \frac{[PL]}{[P][L]} = \frac{1}{K_D}\,. \tag{1.3}$$

K_D can also be computationally estimated through its Gibbs free energy of binding.

$$\Delta G = k_B T ln \frac{K_D}{C_0}\,, \tag{1.4}$$

where ΔG is the free energy change upon binding, k_B is the Boltzmann constant (1.3806485210^{23} J/K), T is the temperature and C_0 is the standard state concentration (usually 1 M).

Thermodynamic quantities such as ΔG can be interpreted from an atomistic point of view using the ensemble idea initially developed by Boltzmann and Gibbs.[30],[31] The idea is that every thermodynamic property of a macroscopic system can be calculated as an average from the mechanical property arising in each ensemble.[32],[33] An ensemble is the collection of all possible microstates for N particles under specified thermodynamic conditions. A microstate is a set of configurations (positions and momenta) that describe the position and velocity of each particle. The entire set of positions (q) and momenta (p) of all the particles of a given system is called *phase space*. The imaginary curve in the phase space formed by an entire collection of particles adopting a particular conformation through time is called trajectory. Rather than focusing on the time evolution of the trajectory, an average ensemble property can be computed by means of a distribution function *P*(Γ*)* that describes the ensemble behavior. *P*(Γ*)* is the probability that at some specific time, under specific thermodynamic conditions, the system is in a particular microstate or has a particular energy. So, the computation of the property average can be implemented by calculating the value of the property periodically at times t. This can be done because, based on the "ergodic hypothesis", the ensemble average and the time average should be the same, over infinite period of time.[34] Therefore, the target of statistical mechanics is to determine the probability function, *P*(Γ*)*, in order to retrieve the macroscopic thermodynamic quantities of a system.

In order to describe and compute the distribution function of an ensemble, it is important to define the type of the ensemble based on the specified thermodynamic conditions. The canonical ensemble (NVT) is the set of all possible positions and momenta for all the particles such that the number of particles (N), the system volume (V) and the temperature (T) are constant and it specifies variation of energy. It describes the possible states of the system that is in thermal equilibrium with a heat bath and it is very useful for single-phase properties at fixed density. It is not very applicable at phase transitions or structural changes that involve a change in volume, for instance freezing or boiling. In this ensemble the probability function, *P*(Γ*)*, is described by the Boltzmann distribution. Thus, the probability of a particle to be in state *i*, defined by a specific set of position q^N and momentum p^N and the total energy described by the Hamiltonian function $H_i(q^N, p^N)$ is:

$$P_i(\Gamma) = \frac{e^{-\beta H_i(q^N, p^N)}}{\sum_i e^{-\beta H_i(q^N, p^N)}} \tag{1.5}$$

where $\beta = \frac{1}{k_B T}$. The denominator of the Boltzmann distribution is called the partition function and is often indicated as Q_{NVT}. It calculates the number of microstates accessible to the system at a particular temperature through the following equation.

$$Q_{NVT} = \sum_i e^{-\beta H_i(q^N,p^N)} . \tag{1.6}$$

In classical statistical mechanics, the set of microstates is uncountable, so the partition function is expressed as an integral rather as a sum.

$$Q_{NVT} = \frac{1}{h^{3N}} \frac{1}{N!} \int e^{-\beta H(q^N,p^N)} dq^N dp^N , \tag{1.7}$$

where the $\frac{1}{h^{3N}}$ term is used to make the quantity dimensionless (h is the Plank's constant, 6.6210^{-34} J) and the $\frac{1}{N!}$ term takes into account that the N particles are indistinguishable. If we assume that the kinetic $K(p^N)$ and potential energy components $U(q^N)$ of the Hamiltonian function are separable then from Equation 1.7 we have:

$$Q_{NVT} = \frac{1}{h^{3N}} \frac{1}{N!} \int e^{-\beta U(q^N)} dq^N \int e^{-\beta K(p^N)} dp^N = Q_{id} Z_{NVT} . \tag{1.8}$$

Q_{id} is the momentum integral that can be analytically obtained by:

$$Q_{id} = \frac{V^N}{\Lambda^{3N} N!}, \tag{1.9}$$

where Λ is the thermal de Broglie wavelength and is given by Equation 1.10:

$$\Lambda = h^2 / (2\pi m k_B T)^{1/2} , \tag{1.10}$$

where m is the molecular mass. Z_{NVT} is the configuration integral and it usually cannot be evaluated analytically:

$$Z_{NVT} = \int e^{-\beta U(q^N)} dq^N . \tag{1.11}$$

The partition function Q_{NVT} plays a vital role in statistical mechanics, as it is linked with the Helmholtz free energy, A_{NVT}:

$$A_{NVT} = -k_B T ln(Q_{NVT}) . \tag{1.12}$$

The passage from the Helmholtz free energy to the Gibbs free energy, G_{NPT} described in Equation 1.4 is done through translating the canonical ensemble concepts into the isothermal-isobaric ensemble (NPT). In this ensemble, the number of particles (N), the pressure (P) and the temperature (T) are fixed. Laboratory experiments are typically executed under these specific thermodynamic conditions. The partition function, Δ_{NPT} can be written as:

$$\Delta_{NPT} = \int\int\int e^{-\beta H(q^N,p^N)} e^{-\beta pV} V^N dq^N dp^N dV \,. \tag{1.13}$$

From this equation it is possible to prove that Δ_{NPT} is related to G_{NPT} through:

$$G_{NPT} = -k_B T ln(\Delta_{NPT}) \,. \tag{1.14}$$

Using the same approach as in the NVT ensemble, we can derive that the configurational integral, Z_{NPT},

$$Z_{NPT} = \int\int e^{-\beta U(q^N)+pV} dq^N dV \,. \tag{1.15}$$

Therefore, G_{NPT} can be calculated from Z_{NPT} that can in turn be computed by determining the potential energy of the system. A good technique to solve this problem numerically is Molecular Dynamics (MD) simulations, that allows the sampling of the potential energy surface of the molecular system.

1.3 Molecular Dynamics

1.3.1 Integrators

The Molecular Dynamics simulation method is based on Newton's second law or the equation of motion that dictates how atoms move subject to their interactions. Newton's second law is written as:

$$\mathbf{F} = m\mathbf{a} = m\frac{dv}{dt} = \frac{dp}{dt} = m\frac{d^2r}{dt^2} = -\frac{dU}{dr_i} \,, \tag{1.16}$$

where $\boldsymbol{F}$ is the force that acts on the particle, m is the mass of the particle, $\boldsymbol{a}$ is the acceleration, v is the velocity, p is the momentum, r is the position of each particle and U the potential energy function of the system. A trajectory is produced by integrating these equations of motion and describes the positions, momenta and accelerations of the particles. Once the positions and momenta of each particle are known, it is a straightforward process to calculate where the particles should move at any given time. It is worth mentioning, that energy and momentum are conserved by the equations of motion (see Appendix A).

There are many different integrators to solve numerically the equations of motion, but in the context of this thesis the velocity Verlet algorithm was used. The basic idea behind this algorithm is to use the information about the positions and the velocities at time t, to predict where the particles would be in a small time in the future, t+δt. This may be done using a Taylor series expansion, as is expressed by the following equation:

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \mathcal{O}\delta t^3\,, \tag{1.17}$$

where $\mathcal{O}\delta t^3$ denotes all the Taylor series terms with order greater than 2. Forces depend only on the positions of the particles, so the acceleration, $a(t)$ can be computed once position and velocities are known. The prediction of the velocities at time t+δt is implemented by using the acceleration at time t and correcting the estimation by using the acceleration at one time step at the future, $a(t$+$\delta t)$, through Equation 1.18:

$$v(t+\delta t) = v(t) + \frac{1}{2}[a(t) + a(t+\delta t)]\delta t + \mathcal{O}\delta t^3\,. \tag{1.18}$$

In practice, velocity-Verlet algorithm splits Equations 1.17, 1.18 in three parts:

$$v(t+\frac{\delta t}{2}) = v(t) + \frac{1}{2}a(t)\delta t\,, \tag{1.19}$$

$$r(t+\delta t) = r(t) + r(t+\frac{\delta t}{2})\delta t\,, \tag{1.20}$$

$$v(t+\delta t) = v(t+\frac{\delta t}{2})\delta t + \frac{1}{2}a(t+\delta t)\delta t\,, \tag{1.21}$$

where the first equation computes the velocities at time $t+\frac{\delta t}{2}$, using the velocities and the accelerations from time t. This allows the calculation of the positions at time t+δt and finally the update of the velocities at time t+δt.

1.3.2 Force-Fields

In principle, the force on each atom, and thus the interactions that dictate its movement, could be calculated at each MD time step using the fundamental equation, Schrödinger equation, of the quantum mechanics. One has to take the positions of the atoms, solve the Schrödinger equation and calculate the forces exerting on each atom, advance the position of the atoms by a small time-step and then recalculate the forces of the atoms. This is currently computationally very expensive and limits the number of atoms one can simulate to ca. 100 atoms for a brief (ca. ps) amount of time. Therefore, a simpler approach is necessary to simulate complex systems over longer timescales. A useful assumption is the Born-Oppenheimer approximation. This approximation enables the electronic and nuclear wavefunctions to be separated, as the electrons move much faster than the nuclei.[35],[36] It also allows the description of a nucleus, i, as a simple classical mechanics system, by making use of the Schrodinger equation for the nuclei interactions:

$$\frac{1}{2}m_i(\frac{\partial r_i^2}{\partial t}) + U_i(r_i) = E_{tot}(r_i)\,, \tag{1.22}$$

where r is the position of the nuclei, m is the mass of the nuclei, $\frac{1}{2}m_i(\frac{\partial r_i^2}{\partial t})$ is the kinetic energy of the nucleus and $U_i(r_i)$ is the nucleus-nucleus interaction potential energy function.

As a consequence, the interactions and the forces between atoms are described as a function of the nucleus positions only. For this purpose, simple mathematical functions with a set of optimized parameters are used and they are called force-fields. These functions are rapid to evaluate, and can be broken down into various components:

1. **Non-bonded interactions (repulsions and Van der Waals attractions):** Is an interaction evaluated between certain pairs of non-bonded atoms. The potential is positive when the particles are close to each other, because they repel each other. This is because of Pauli exclusion principle, which causes the energy of the system to increase rapidly as the separation decreases. This repulsion is responsible for the hardness of the materials as it stops atoms from overlapping with each other. When the particles are separated by longer distances, they attract each other with dispersion interactions. This is a quite rapidly decaying attraction as it dies off with the distance raised to its sixth power.

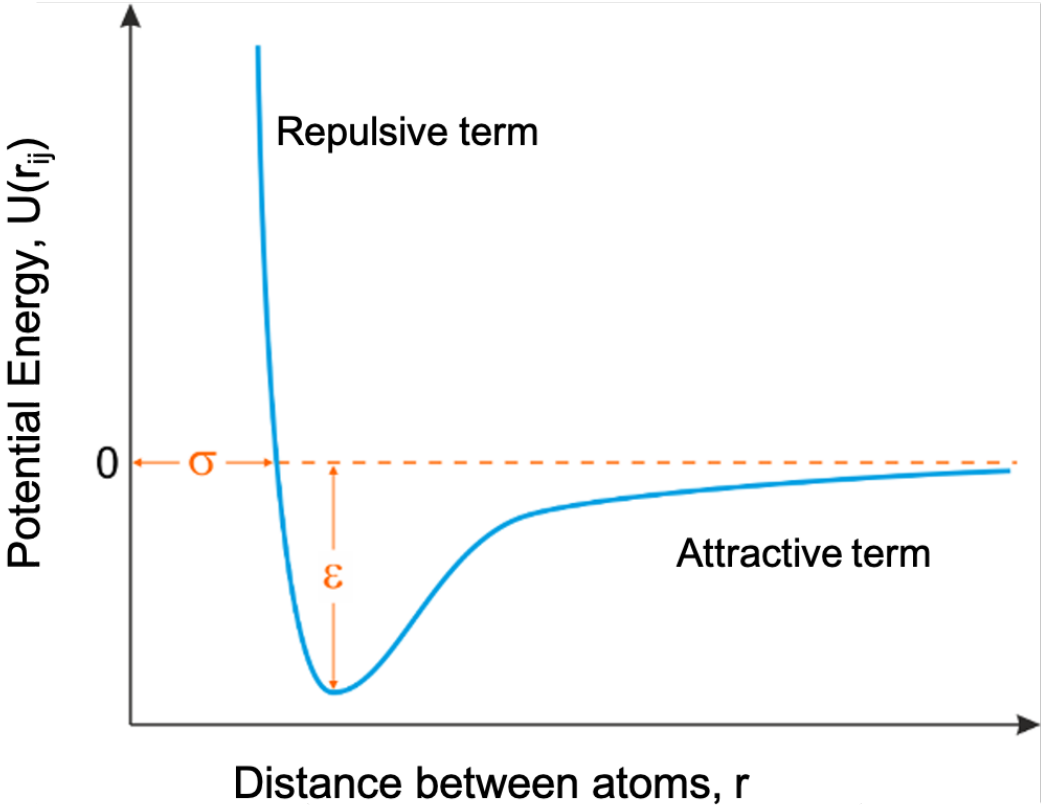


FIGURE 1.3: The Lennard Jones potential curve, where the atoms repel each other at short distances due to the repulsive term and attract each other at longer distances due to the attractive term.[37]

The Lennard-Jones potential (Figure 1.3) is a convenient description to represent the combination of attractions and repulsions, is given by the following Equation:

$$U_{LJ}(r_{ij}) = 4\epsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}], \tag{1.23}$$

where ϵ_{ij} is the Lennard-Jones well depth, which is the minimum energy that the particles experience. It is an energy parameter and indicates the strength of the interaction. σ_{ij} is the effective diameter of the atoms and it is a distance parameter that determines the size of the atoms. The second term is the attraction term and is derived from quantum mechanical calculations. The first term is the repulsion term and based on quantum mechanical calculations, it should have an exponential dependence on interatomic distance. However, the calculation of the exponential function is five times more expensive in computational time than the simple mathematical functions for instance multiplication. Therefore, the computation of r^{12} is implemented by the multiplication of r^6 by itself, which is a very cheap operation.

These numbers are evaluated for two identical atoms of the same type. The interaction energy between two dissimilar non-bonded atoms is provided by a series of equation called combining rules. The most widely used are the Lorentz-Berthelot mixing rules[38], that are provided by Equations 1.24, 1.25:

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}} \tag{1.24}$$

$$\sigma_{ij} = \frac{\sigma_{ii} + \sigma_{jj}}{2}, \tag{1.25}$$

where ϵ_{ii} and σ_{ii} is the well depth and the effective diameter between two similar non-bonded atoms respectively, and ϵ_{ij} and σ_{ij} is the well depth and the effective diameter between two dissimilar non-bonded atoms.

2. **Non-bonded interactions (electrostatics):** Two non-bonded atoms with a charge or partial charge can interact electrostatically with each other via Coulomb's law as illustrated with Equation 1.26:

$$U_C(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}, \tag{1.26}$$

where q_i, q_j are the charges on the atoms, ϵ_0 = 8.854 x 10^{-12} Fm^{-1} is the vacuum dielectric permittivity and r_{ij} is the distance between the atoms. There are three forms that are followed, based on the modelling purposes:

   (a) In the simplest case the atomic charge is a fixed parameter, therefore atoms carry their assigned charges in every situation. if the force-field is designed to study a particular molecule, partial charges are used to reproduce accurately an experimental or computational electrostatic observable of the molecule.

   (b) Alternatively, the charges are determined from a scheme that depends on the electron negativity of the atoms involved in the interaction. It is useful in force-fields with reduced number of atom types as it maintains flexibility in the recognition of different chemical environments. This flexibility is crucial for the charge, because the electrostatic energy can be very large compared to other components of the force-field.

   (c) Finally, account for the induced polarisation when treating electrostatic interactions. The strength of the Coulombic interactions is dependent on the dielectric constant, ϵ_r, of the intervening medium. The dielectric permittivity affects the decay of the electrostatic interactions by *1/*ϵ_r. In addition, it changes the electrostatic response of atoms or molecules to the presence of a charge: smaller dielectric constants correspond to smaller responses to the presence of a nearby charge. Polarisation plays a vital role for describing key biomolecular interactions such as cation–π interactions. It is also important to describe the change of the polarisation of molecules when they encounter different interacting partners during the course of a simulation. For instance when a solute is moving from a non-polar region of the system to a polar region, the polarisation is increased. However, the limiting factor of using such approach is the increased computational cost.[39]

Hydrogen bonding interactions are also described via electrostatics as this is a polar interaction, which is formed between heteroatoms and hydrogen atoms that are not formally bonded.

3. **Bonded interactions (bond stretching potential):** This term is applied to pair of atoms that share a covalent bond. The energy of a bond is at its lowest value at a reference length, called the equilibrium length. If the bond is compressed, with respect to the equilibrium length, the electron clouds of the two atoms will progressively overlap and the energy will rapidly increase. Alternatively, if the bond is stretched beyond equilibrium the energy starts to increase up to a point where the bond disassociates. The energy, $U_{stretch}$, can be expressed by taking a Taylor expansion about the equilibrium energy distance, r_0, as it is illustrated in Equation 1.27:

$$U(stretch) = U(r_0) + \frac{dU}{dr}|_{r=r_0} + \frac{1}{2!}\frac{d^2U}{dr^2}|_{r=r_0}(r-r_0)^2 + \frac{1}{3!}\frac{d^3U}{dr^3}|_{r=r_0}(r-r_0)^3 + ... \tag{1.27}$$

If we assume that the first two terms are zero, the first by arbitrary choice and the second because r_0 is the minimum of the function, we can obtain Equation 1.28:

$$U(stretch) = \frac{1}{2}k_r(r-r_0)^2, \tag{1.28}$$

where k_r is the force constant and represents the stiffness of the spring. Chemical bonds are assumed to be classical harmonic springs and are parameterised with the equilibrium length, that can be obtained from X-ray structural data and with the force constant, which can be acquired by stretching frequencies in real molecules. The real potential is asymmetrical, so we approximate the bottom of the potential well with a harmonic function. For strongly bound pairs of atoms, the harmonic potential is very good. However, as any truncated Taylor expansion, this potential works better in regions near its reference point, r_0. Thus, for weekly bound molecules, where the bond is stretched to longer r values, the energy keeps rising quadratically.

4. **Bonded interactions (bond bending potential):** Additional functions are used to enforce reasonable molecular geometries between groups of three atoms that share two consecutive bonds. For instance if we have a sp^3 carbon, it would be close to the tetrahedral angle (109°). For the bond bending potential, a harmonic function is also used through Equation 1.29:

$$U(bend) = \frac{1}{2}k_\theta(\theta - \theta_0)^2, \tag{1.29}$$

where θ_0 is the equilibrium bond angle, that may be obtained from X-ray structural data and k_θ is the force constant that is acquired from bonding frequencies in real molecules. The energy needed to stretch an angle away from the equilibrium bond angle is much lower than the energy needed to distort a bond, so bond bending force constants tend to be smaller than the bond stretching ones. Finally, the accuracy can be improved by adding higher order terms.

5. **Bonded interactions (proper torsional potential):** The torsional potential is a function of a dihedral angle, which is the angle between four atoms. As this is a periodic function, it is conveniently expressed by a Fourier series, as shown in Equation 1.30:

$$U(torsion) = \sum_{n} \frac{1}{2} V_n [1 + (-1)^{n+1} cos(n\phi - \gamma)] \,, \tag{1.30}$$

where n is the set of periodicities, V_n are the torsional rotation force constants, ϕ is the current torsional angle, and γ are the phase angles, which are usually chosen to define where the torsion angle passes its minimum value. The factor of $\frac{1}{2}$ is included so that the term amplitude of V_n is equal to the maximum the particular term can contribute to U. The factor of *(-1)*$^{j+1}$ so that the term in brackets is zero for all n when ϕ is equal to π. The number of terms needed in the Fourier series depends on the complexity of the torsional potential and the desired accuracy. For organic compounds, three terms are generally used and they are depicted in Figure 1.4.

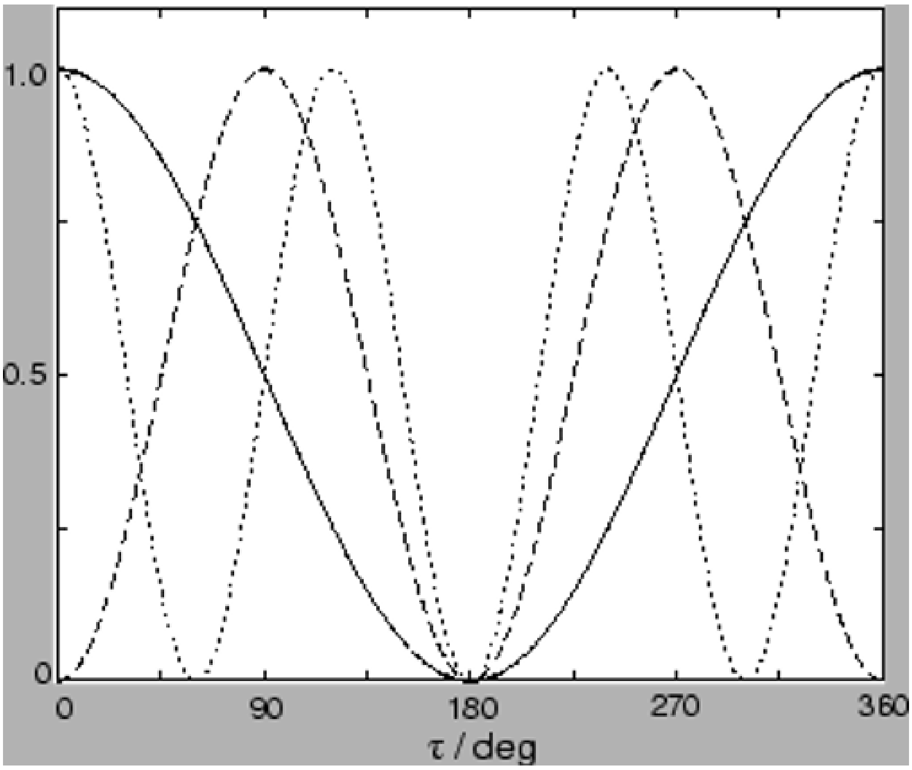


FIGURE 1.4: The proper torsional potential where the bold line refers to n=1, the dashed line to n=2 and the dotted line to n=3 Fourier components. Adapted from `http://cmt.dur.ac.uk/sjc/thesis_dlc/node74.html` [40]

The various parameters can be derived from *ab-initio* quantum calculations, but this has to be done in conjunction with other non-bonded and bonded parameters. This is because the total energy of the molecule does not only depend on the torsional potential, but also on the non-bonded interactions.

6. **Bonded interactions (improper torsional potential):** This is a constraint potential. If we take four atoms (A, B, C, D) that form three angles (A-B-C, A-B-D, and C-B-D) then B must be retained in the same plane formed by A, C, and D. Therefore, a constraint must be issued through the definition of a potential, which is a function of an angle ϕ, between two planes. In this case, the two planes are A-B-C and D-B-C and the potential is harmonic, as it is shown in Equation 1.31:

$$U(constraint) = \frac{1}{2}k_{\phi}(\phi - \phi_0)^2\,, \tag{1.31}$$

where ϕ_0 is the equilibrium value that is determined by quantum mechanics. This potential assures that ϕ angle remains near to its equilibrium value, thus that the chemical bonds stay in the right geometry.

1.3.3 Treatment of the Interactions in Molecular Dynamics

One of the limitations of the computer simulations, is the number of atoms that one can simulate. The computational cost limits the number of atoms and the number of interactions with other molecules, so MD simulations normally include 10^2-10^6 atoms in a box. The problem that occurs is that the behaviour of finite size systems is different from that of systems used in experiments. The main difference is that in a MD simulation, many atoms are close to the walls of the box, so surface effects may influence the system properties. On the other hand, in a real macroscopic system, surface effects do not play a vital role in the simulation.

A method to avoid such artifacts is the so-called Periodic Boundary Conditions (PBC).[41] In this approach, the walls are removed and the bulk material is assumed to be made up of periodic arrays of replicas of a central box. In this way, when a particle enters or leaves the boundaries of a simulation cell, an image particle simultaneously leaves or enters the simulation region from the point related to the entrance or exit location by lattice symmetry. Therefore, the atoms are allowed to leave the box, are not influenced from surface effects and are always replaced by replicas. An example of this method is illustrated in Figure 1.5

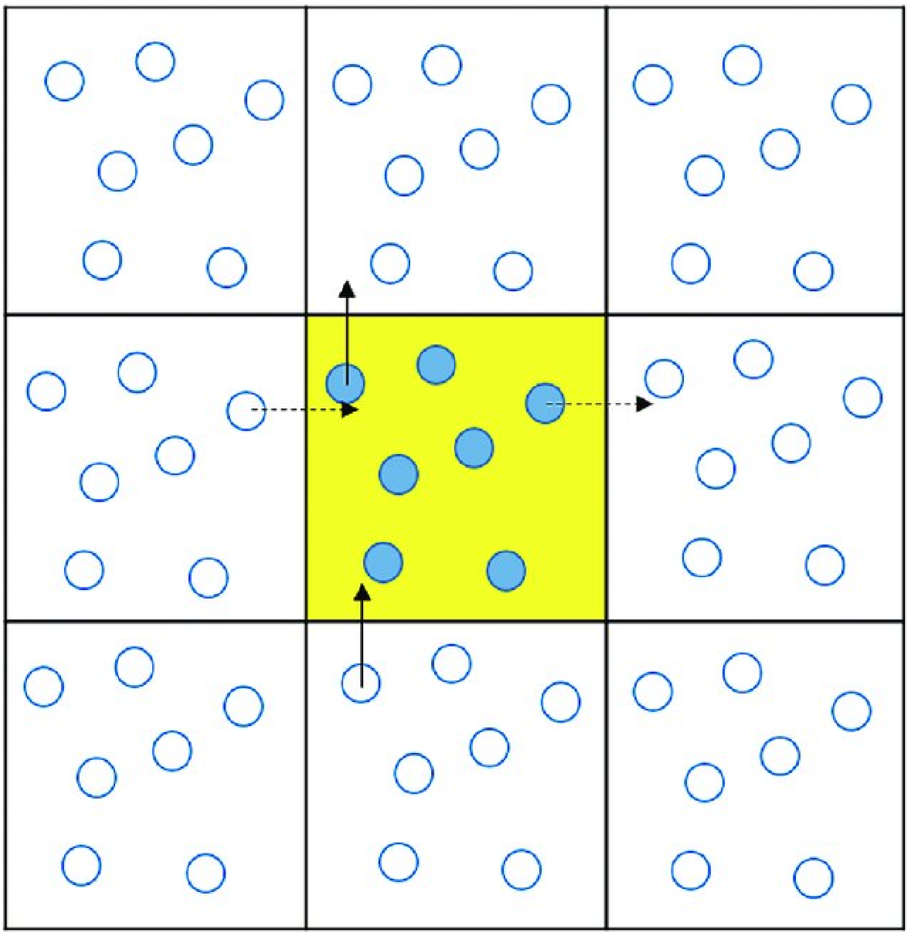


FIGURE 1.5: The periodic boundary conditions, where the central simulation box is coloured with yellow. Filled blue circles represent particles inside the central simulation box, while open circles represent their images in other periodic cells. Dashed and bold lines shows movement of two particles near the boundaries. As a particle leaves the central simulation box, its periodic image enters the box from the opposite end. Adapted from Katiyar's book titled "*Molecular simulations in drug delivery: Opportunities and challenges*".[42]

PBC ensures the conservation of mass, total number of molecules and total energy in the simulation cell.

As a result of using PBC, each particle can interact not only with the other atoms in the simulation cell, but also with all the other replicas. For short range interactions, the problem can be solved by picking a finite range potential within the criteria of minimum image convention. Therefore, we take into account only the strongest interaction, which is the one with the shortest distance between the neighbouring particles. It is not necessary that the interaction will be between atoms in the same box.

In practice, the mechanism of implementing this is to truncate the potential in a finite range and assume that it is zero beyond some finite length called cut-off distance, r_c. The maximum cut-off distance must be equal or less than the half of the simulation box. It is also a good idea, to shift the potential, so that the interaction energy is zero at the cut-off. The point of that is that the force is the derivative of the energy, so if the potential is not shifted, the force will be discontinuous near the cut-off, which may cause a numerical instability in the integration of the equations of motion.

For electrostatic interactions, the minimum image convention is insufficient. The range of these interactions is much longer than the size of the simulation box that can be considered in a MD simulation. Thus, it is inappropriate to truncate the potential as the Coulomb interactions have very large effects over a very long distance. This problem is solved with techniques that are more expensive than the simple truncation, but they respect the long-range character of the forces.

The Ewald method[43] is the most widely used for computing the long-range contributions to the potential energy (see Appendix A.2). An alternative approach to Ewald summation is the reaction field method, which is faster but less accurate.[44] This technique works in a manner similar to the simple truncation method. A "cavity" α with a cut-off sphere of r_c is defined and the region outside that cavity is assumed to be a dielectric continuum with a dielectric constant ϵ_{RF}. The particles in the ensemble polarize the surrounding dielectric constant and this produces an electric field E_α represented by Equation 1.32:

$$E_a = \frac{2(\epsilon_{RF}-1)}{2(\epsilon_{RF}+1)}\frac{1}{r_c^3}\sum_b \mu_b\,, \tag{1.32}$$

where the summation is over all molecules in the cavity of molecule α and

$$\mu_b = \sum_{i \in Mol_b} q_i r_i\,, \tag{1.33}$$

is the dipole moment of a molecule b. In addition, the effective pairwise potential becomes:

$$U(Coulomb) = q_i q_j [\frac{1}{r_{ij}} + \frac{(\epsilon_{RF}-1)r_{ij}^2}{2(\epsilon_{RF}+1)r_c^3}]\,. \tag{1.34}$$

1.3.4 Implementation of Molecular Dynamics

For the implementation of a MD simulation, a set of initial configurations are required for the beginning of the process. The final results should not be affected by the selection of the initial positions and velocities. Initial velocities are typically drawn randomly from the Maxwell-Boltzmann distribution:

$$p(\mathbf{V}_i) = (\frac{m_i}{2\pi k_B T})^{\frac{1}{2}} e^{-\frac{1}{2}\frac{m_i \mathbf{V}_i^2}{k_B T}}\,. \tag{1.35}$$

The above equation calculates the probability that an atom i with mass m_i at temperature T has a velocity vector $\boldsymbol{V}_i$ = (V_{ix}, V_{iy}, V_{iz}). Ultimately, the system may require minimization, in order to remove any artificial structure and heating to reach the desired temperature of the simulation.

After the assignment of the initial configurations, a period of equilibration should be implemented under the desired conditions of pressure, temperature, etc. It is an essential procedure, as it permits the monitoring of the system and assures that everything works correctly. Once the system is equilibrated, the observer starts to measure the properties of the system that they want to evaluate. This is called the production run and it allows one to produce results after discarding the data from the equilibrium run.[45]

One important aspect of the MD simulations is the choice of the time-step.[46] A large enough time-step is needed to simulate a long enough real time, but also it should be small enough to satisfy the conversation laws of total energy and momentum in the integration of the equations of motion. Typically, relative errors in the total energy or momentum up to 1x10^{-4} kcal/mol are acceptable.

In chemical applications, the time-steps used are always the same and they are shorter than the fastest motion of the examined system. In molecules, the vibrational moves between bonds are faster than the rotational or translational moves. So, the time-step should be shorter compared to the fastest molecular vibrational period. From IR spectroscopy, the shortest vibrational frequencies arise from bonds between light atoms (H) and heavy atoms (C or O) and they range from 2800-4000 cm^{-1}. If this is converted to vibrational period, it equals to τ = *8.3x10*$^{-15}$ – *1.2x10*$^{-14}$ *s*. Thus, in MD simulations, time-steps of 1 femtosecond (10^{-15} s) are chosen, in order to be 10 times shorter of the molecular vibrational period. This gives a good balance between conserving the total energy and momentum and being long enough for reasonable real time scales up to nanoseconds.

Another essential parameter in MD is to find a way to keep constant the temperature and the pressure, as in real experiments. The technique used to conserve temperature in this work is called Andersen thermostat.[47] In this approach, the system is coupled to a heat bath that establishes the desired temperature. The coupling is represented by stochastic impulsive forces that act occasionally on random particles. The coupling strength is regulated by the frequency of stochastic collisions, *v*. If the successive collisions are assumed to be uncorrelated, then the distribution of time intervals between two successive collisions, *P(t;v)* is of the Poisson form. So, the probability that the next collision will happen in the interval *[t, t+dt]* is expressed in Equation 1.36 as:

$$P(t;v) = ve^{-vt}\,. \tag{1.36}$$

After the collision with the heat bath, the particles are assigned by new velocities that are drawn from the appropriate Maxwell-Boltzmann distribution for the desired temperature. Andersen thermostat generates good results for time-independent properties such as the equation of state of a system or the potential energy. However, it is not an appropriate method for time-dependent properties as the diffusion coefficient, as the dynamics produced by this approach are nonphysical. The stochastic collisions change the dynamics in a way that is not real as they lead to sudden random decorrelation of particle velocities.

Regarding the pressure, there are also many barostats that keeps it constant, but in this report the Monte Carlo barostat is used.[48] In this approach, after using the velocity Verlet algorithm for a time-step δt, a Monte Carlo move is made by adding or subtracting a random increase to the volume of the system. For cubic boxes, the new volume is determined by Equation 1.37:

$$V' = V + R[S(\delta V)]\,, \tag{1.37}$$

where R is a random number between -0.5 to 0.5, δV reduces the maximum size of the volume increment and S is an adjustable scaling factor. The change in volume is transferred to the particles positions by scaling the coordinates in Equation 1.38:

$$r'(t+\delta t) = r(t+\delta t)[\frac{v'^{\frac{1}{3}}}{v^{\frac{1}{3}}}]\,. \tag{1.38}$$

The box move is then accepted or rejected using the Metropolis algorithm with the sampling Equation 1.39:

$$\Delta W = (E'-E) + P_0(V'-V) - Nk_BT_0ln\frac{V'}{V}\,, \tag{1.39}$$

where E' is the new energy, P_0 and T_0 are the external pressure and temperature and N is the number of particles of the system. The probability which the box moves are accepted is expressed in Equation 1.40 as:

$$P(\Delta V) = \begin{cases} e^{-\frac{\Delta W}{k_B T_0}}, & \Delta W > 0\,. \\ 1, & \Delta W \leq 0\,. \end{cases} \tag{1.40}$$

A successful move is completed by updating the forces on the particles to generate a set of positions and accelerations for the new configuration. Alternatively, if the move is rejected, the original configuration is restored.

1.4 Free Energy Calculations

1.4.1 Alchemical Free Energy Methods

According to Equation 1.13, the Gibbs Free Energy, G_{NPT}, can be computed using the configuration integral, Z_{NPT}. The direct calculation of Z_{NPT} is numerically impossible, due to the high dimensionality of protein-ligand complexes. Instead, it is more tractable to compute ratios of the configuration integrals between two related thermodynamic states, A and B. This observation forms the basis of free energy calculation techniques.

Two major methods can be used to calculate the free energy differences between thermodynamic states. The first approach is called Free Energy Pertubation (FEP), introduced by Zwanzig in 1955.[49] As mentioned previously, MD simulations offer a good approach to evaluate Z_{NPT}. If one implements MD simulations using the potential energy function of thermodynamic state A, then Equation 1.41 may be used to compute the free energy change of replacing A with B:

$$\Delta G^{EXP}(A \rightarrow B) = -k_B T ln < e^{-\beta[U_B(q) - U_A(q)]} >_A\,, \tag{1.41}$$

where the angular brackets <> indicate that the quantity inside is averaged over all the configurations of A and weighted by their Boltzmann probabilities. The procedure involves periodically computing the potential energy that B will have for a given q_i value and subtracting this from the potential energy of A at the same q_i. Ultimately, the free energy difference is calculated by using the evaluated potential energies. Therefore, the free energy between two thermodynamic states A and B is the Boltzmann weighted probability of the difference of the potential energies between A and B. It is also possible to perform the reverse process as illustrated in Equation 1.42:

$$\Delta G^{EXP}(B \rightarrow A) = -k_B T ln < e^{-\beta[U_A(q) - U_B(q)]} >_B \ . \tag{1.42}$$

If it is assumed that the number of samples is infinite, the free energy changes between the two processes must be equal. However, this is not the case in practice because datasets are necessarily finite. In order to calculate the deviations from the expected results the quantity h is used, which is called the hysteresis of the results. It may be defined as the absolute value of the sum of the two free energy changes and it is expressed in Equation 1.43:

$$h = |\Delta G^{EXP}(B \rightarrow A) + \Delta G^{EXP}(A \rightarrow B)| \, , \tag{1.43}$$

where h should have a value as low as possible, as it is an indicator of the consistency of the results. A now deprecated approach consisted in calculating the hysteresis in both directions. The problem with this strategy is the asymmetry in the rate of convergence of the free energy estimate to the true free energy change. This asymmetry can be understood in terms of state space overlap between the low energy configurations of A and B. Ideally the reference state should be the state of higher entropy, as the low energy configurations of the perturbed state is more likely to be a subset of the low energy configurations of this state. Unfortunately, it is difficult to know *a priori*, which protein-ligand complex has the higher entropy, therefore it is not easy to determine beforehand in which direction the FEP equation converge more rapidly.

A usual way to deal with this problem is by multi-staging the transformation of A to B. A coupling parameter λ is defined to control this conversion, thus the direct transformation of A to B is separated in k intermediate steps such that λ_0 = A = 0.0 and λ_1 = B = 1.0 as illustrated in Equation 1.44:

$$\Delta G^{EXP}(A \rightarrow B) = \sum_{k=0}^{k=n-1} \Delta G^{EXP}(\lambda_k \rightarrow \lambda_{k+1})\,. \tag{1.44}$$

The result is that the exponential averaging is only implemented between states that have high degree of phase space overlap. Although the number of simulations is increased by a factor k, each of these simulations converge faster and the overall process is more accurate. However, due to the fact that the intermediate states are unphysical and not of any interest, it would be useful to minimize their number in order to minimize the computational cost.

There are number of ways to implement the aforementioned strategy. The most efficient and widely used technique is the Bennet Acceptance Ratio (BAR) method:

$$\Delta G^{BAR}(A \rightarrow B) = -\beta^{-1} ln \frac{< f(\beta[U_{A(q)} - U_{B(q)} - C] >_B}{< f(\beta[U_{B(q)} - U_{A(q)} - C] >_A} + C\,, \tag{1.45}$$

where the numerator of the ratios is the ensemble average of the function *f*. It takes as input β times the difference between U_A and U_B for a given microstate q minus a constant C. This constant is obtained from data sampled from equilibrium distribution *f(B)*. The denominator is the ensemble average of the same function but with opposite sign. For this equation is necessary to assume that the same number of samples is used for both datasets. For a finite number of samples, the statistical optimal choice that minimizes the standard error is provided by the Fermi function as shown in Equation 1.46:

$$f(x) = \frac{1}{1+e^x} \quad and \quad C = \Delta G\,. \tag{1.46}$$

Since ΔG is usually not known in advance the equation must be solved self-consistently. Firstly, a guess is made for C and then the ratios from Equation 1.42 are solved to obtain ΔG and then one sets $C = \Delta G$. The process is iterated until ΔG is not changing anymore. It can be proved that this procedure always converges to the most accurate ΔG given available data. It is also possible to generalize BAR to handle multiple thermodynamic states at the same time and this can slightly improve the precision of the free energy estimate. This method is called multistate Bennet acceptance ratio (MBAR).[50] In addition, this approach can be used to evaluate energies of different perturbed states from a reference state. The vital part in this calculation is to find the reference state such that all the binding free energies converge well. Occasionally, it is useful to construct a non-physical state in order to maximize the overlap with the set of the perturbed states.

An alternative approach to compute the free energy change between the two states A and B is thermodynamic integration (TI).[51] It is a straightforward application of the fundamental rules of calculus as it expressed in Equation 1.47:

$$\Delta G^{TI}(A \rightarrow B) = \int_{\lambda=0}^{\lambda=1} (\frac{\partial G}{\partial \lambda}) d\lambda \, , \tag{1.47}$$

where A is defined at λ=*0* and B is defined at λ=*1*. This relationship is related to the ensemble average through Equation 1.48:

$$\int_{\lambda=0}^{\lambda=1} (\frac{\partial G}{\partial \lambda}) d\lambda = \int_{\lambda=0}^{\lambda=1} < \frac{\partial U}{\partial \lambda} > d\lambda \, , \tag{1.48}$$

where one can implement TI by deriving an analytical expression for the first derivative of U with respect to λ and evaluate it through MD simulations. Alternatively, a double wide sampling strategy called finite difference thermodynamic integration (FDTI) can be applied to estimate the free energy gradients. So, the free energy change between λ and λ+$\Delta\lambda$ is provided by Equation 1.49:

$$\frac{\partial G}{\partial \lambda} \approx \frac{\Delta G}{\Delta \lambda} = \frac{G(\lambda + \Delta\lambda) - G(\lambda)}{\Delta \lambda} = \frac{\Delta G(\lambda \rightarrow \lambda + \Delta\lambda)}{\Delta \lambda} \tag{1.49}$$

It is essential that $\Delta\lambda$ is sufficiently small, for this equation to be valid. On the other hand, it should not be too small, because the numerical precision can be affected from the floating-point rounding error. Therefore, the free energy difference is evaluated using a perturbation technique and equation 1.46 is computed using numerical integration. FDTI has the advantage of not suffering from the asymmetry in convergence of forwards and backwards calculations. Different calculus techniques are used to estimate the integral over the λ interval. For instance, trapezoidal or Simpson's rule are applied if the data points present small curvature. These methods are simple to perform, but they have underestimation and overestimation errors.

In general, BAR generates more accurate results than the exponential averaging for the same amount of resources but requires post-processing analysis of the results. In addition, TI is also more robust than exponential averaging as it lacks hysteresis issues. However, it suffers from overestimation and underestimation errors when one has finite number of data points.

A further development of the multistage approach is the use of thermodynamic cycle. It relies on the fact that the free energy of a system is a state function, therefore the free energy change for a closed thermodynamic cycle is zero. The use of thermodynamic cycles permits the comparison of the predicted free energy values with experimental values. An example of the thermodynamic cycle used in these calculations is depicted in the following figure:

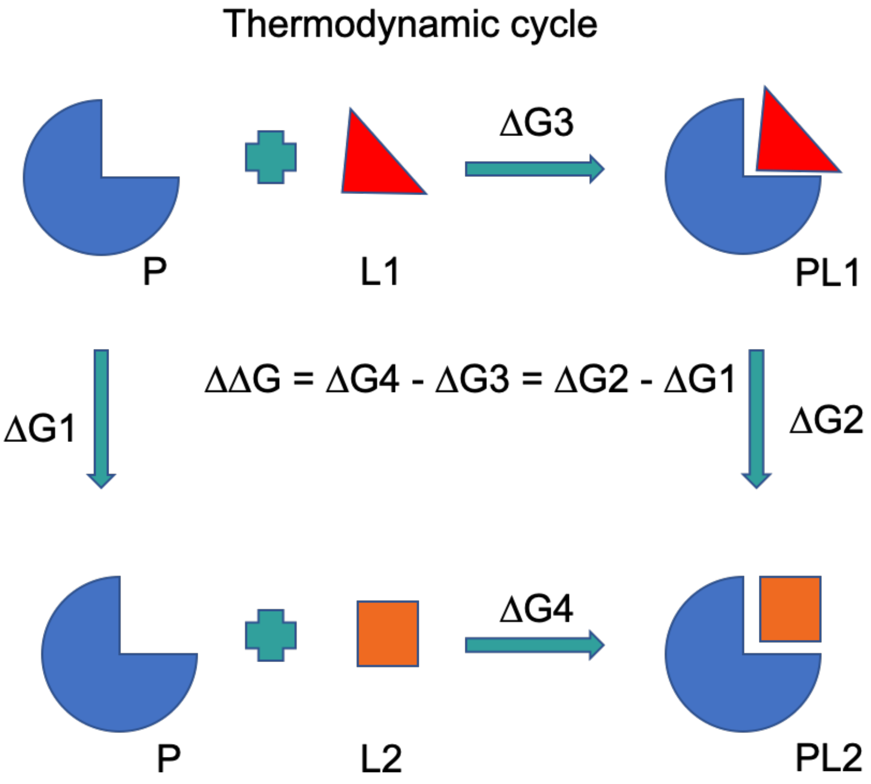


FIGURE 1.6: An example of the thermodynamic cycle used in Alchemical Free Energy (AFE) calculations. The relative binding free energy is depicted from the two horizontal procedures, and is equal to the transmutation of the one ligand to the other, shown by the two vertical processes. Ligand L1 transforms in ligand L2 in both solvent and complex phase using an one-step process.

Much of the research done in this thesis is concerned with the computation of the standard binding free energy of two different ligands *L1* and *L2* to protein *P*. The difference in energy of the two horizontal processes, the relative binding free energy, is equal to the difference of the two vertical processes. These processes correspond to the transformation of *L1* (thermodynamic state A) to *L2* (thermodynamic state B). This process does not need to follow physical principles. The only requirement is that the molecule in the beginning of the procedure is *L1* and the molecule at the end of the process is *L2*. So, the reversible binding of two ligands *L1* and *L2* to a protein *P* can be expressed with Equations 1.50 and 1.51:

$$P + L1 \overset{\Delta G^{\circ}_{bind}(L1)}{\rightleftharpoons} PL1\,, \tag{1.50}$$

$$P + L2 \overset{\Delta G^{\circ}_{bind}(L2)}{\rightleftharpoons} PL2\,, \tag{1.51}$$

where Δ G°_{bind}*(L1)* and Δ G°_{bind}*(L2)* are the standard binding free energies of *L1* and *L2* respectively. Subtraction of Equation 1.50 from Equation 1.51 and rearrangement leads to Equation 1.52:

$$L1 + PL2 \overset{\Delta\Delta G^{\circ}_{bind}}{\rightleftharpoons} PL1 + L2\,. \tag{1.52}$$

The free energy change of this reaction can be measured from Equation 1.53:

$$\Delta\Delta G_{bind} = -k_B T ln K_{eq} = -k_B T ln \frac{[PL1][L2]}{[L1][PL2]} = -k_B T ln \frac{Z_{PL1,solv} Z_{L2,solv}}{Z_{L1,solv} Z_{PL2,solv}}\,, \tag{1.53}$$

where *[PL1]* and *[L1]* represent the concentrations for the complex and ligand 1 respectively and *[PL2]* and *[L2]* represent the concentrations for the complex and ligand 2 respectively, K_{eq} is the equilibrium constant, $Z_{PL1,solv}$, $Z_{PL2,solv}$, $Z_{L1,solv}$ and $Z_{L2,solv}$ are the configuration integrals for host-guest systems for ligand 1 and 2 and the solvent molecules 1 and 2 respectively.

Once a strategy to compute the binding free energy has been decided, it is essential to create a mathematical relationship that allows the smooth convention of the potential energy function of *PL1* into one describing the *PL2*. The method used in this report is called single topology. In its most simple form, it uses the interpolation of the force field parameters of *PL1* and *PL2*. If the only difference between *PL1* and *PL2* is the ligand, then only the ligand parameters are coupled with λ. The coupling can be linear or non-linear, for instance the atomic partial charges are expressed in Equation 1.54:

$$q_{i,\lambda} = \lambda^n q_{PL1} + (1-\lambda)^n q_{PL2}\,. \tag{1.54}$$

Similarly, other force-field parameters such as bond lengths can be coupled with λ. This strategy requires mapping between the initial and end states. A complication arises when the number of atoms in *L1* and *L2* differ, because it requires the potential energies of a number of ligand atoms to be turned off/on during the perturbation. For the creation and deletion of atoms, dummy atoms are usually introduced. These are atoms that do not have charge of Lennard-Jones parameters, thus their non-bonded interaction energy is null. However they remain bonded throughout the perturbation.

A common problem that occurs in this approach is called "end-point catastrophe". This numerical instability usually appears near the final points of the perturbation. It is due to the fact that as the ligand atoms disappear, atomic overlaps between non-bonded atoms are allowed and solvent and protein atoms can occupy the available space from these atoms. To prevent this situation, a soft core potential energy function is introduced as shown in Equation 1.55:

$$U_{nonbonded,\lambda} = (1-\lambda)4e_{ij}[(\frac{\sigma_{ij}^{12}}{(\lambda\delta\sigma_{ij})^6}) - (\frac{\sigma_{ij}^{6}}{(\lambda\delta\sigma_{ij}+r_{ij}^2)^6})] + \frac{(1-\lambda)^n q_i q_j}{4\pi\epsilon_0\sqrt{(\lambda+r_{ij}^2)}}, \tag{1.55}$$

where δ and n are soft core parameters for Lennard-Jones and Coulombic interactions. In this way the overlap between real atoms and dummy atoms does not lead to very large energies, and intramolecular interactions can be controlled.[52]

A special case transformation is when a ligand L is converted into a molecule that is not interacting with the solvent or the protein, as if it were in an ideal thermodynamic state. This process is called double annihilation scheme and it is used for the calculation of absolute binding free energies.[53],[54] An example of the thermodynamic cycle used in these calculations is illustrated in Figure 1.7:

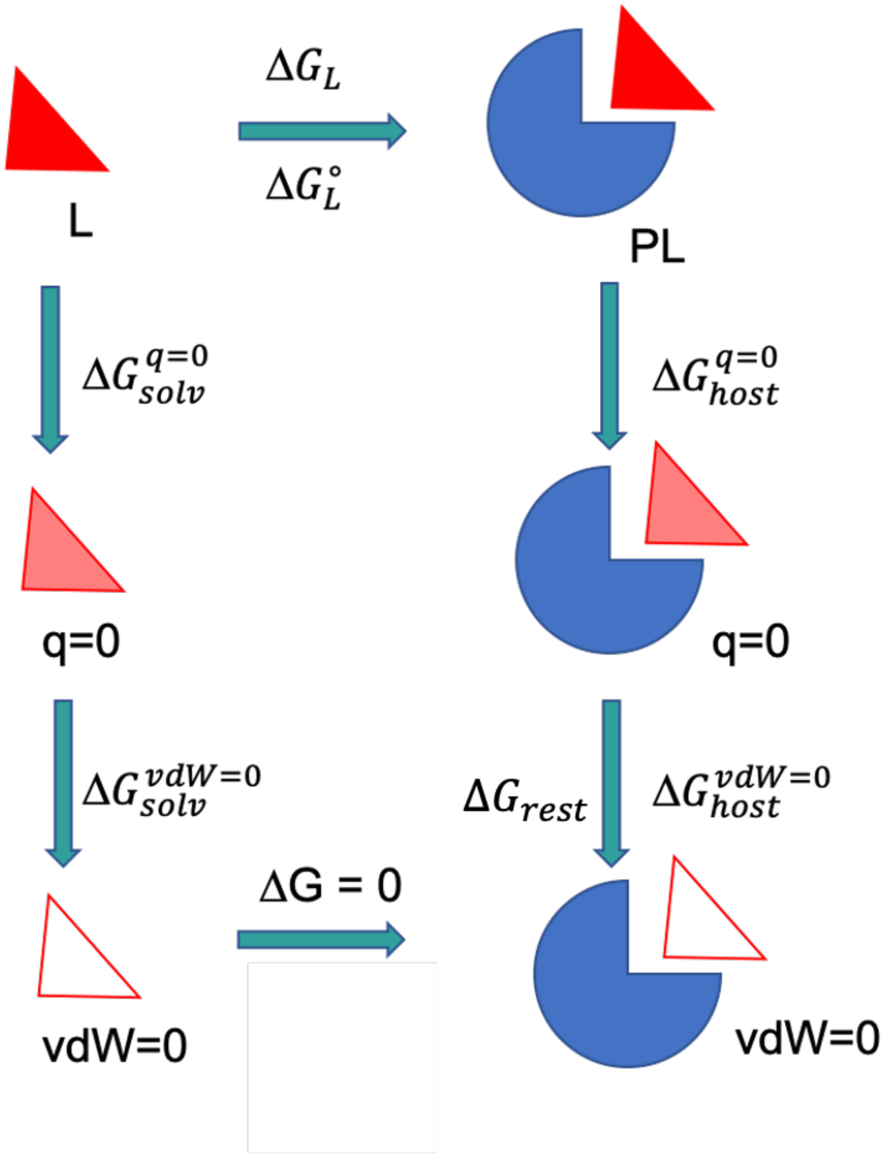


FIGURE 1.7: An example of the thermodynamic cycle used in absolute free energy calculations with the double annihilation method. Ligand L is transformed into an ideal thermodynamic state in both solvated and complex phase using a two-step process. In the *discharging* step, partial charges of the ligand are switched off, retrieving free energy changes $\Delta G^{q=0}_{solv}$ and $\Delta G^{q=0}_{host}$. Subsequently, a *vanishing* step is carried on by turning off the vdW terms of the ligand, providing free energy changes $\Delta G^{vdW=0}_{solv}$ and $\Delta G^{vdW=0}_{host}$. Thus, the absolute free energy ΔG, depicted from the two horizontal processes, can be computed from the difference of the vertical procedures. However, a standard state correction ΔG_{rest} is needed in order to obtain a standard binding free energy ΔG°.

In this method, ligand L is mutated into a "non-interacting" molecule both in the solvated and the bound phase using a two-step process. In the first step, also called *discharging* step, the partial charges of the ligand are turned off giving the free energy changes $\Delta G^{q=0}_{solv}$ and $\Delta G^{q=0}_{host}$. Following, the van der Waals(vdW) terms of the ligand are switched off in the second step (*vanishing* step) providing the free energy changes $\Delta G^{vdW=0}_{solv}$ and $\Delta G^{vdW=0}_{host}$. The final binding free energy ΔG is computed from the two legs of the cycle and can be decomposed into configuration integrals giving:

$$\Delta G = (\Delta G^{q=0}_{host} + \Delta G^{vdW=0}_{host}) - (\Delta G^{q=0}_{solv} + \Delta G^{vdW=0}_{solv}), \tag{1.56}$$

$$\Delta G = -k_B Tln(\frac{Z^{q=0}_L Z^{vdW=0}_L Z_{PL} Z^{q=0}_{PL}}{Z_L Z^{q=0}_L Z^{q=0}_{PL} Z^{vdW=0}_{PL} Z_P}), \tag{1.57}$$

$$\Delta G = -k_B Tln \frac{Z_{PL}}{Z_L Z_P}, \tag{1.58}$$

where $Z^{q=0}_L$ and $Z^{vdW=0}_L$ are the configuration integrals of the ligand in the free phase for the *discharging* and *vanishing* step respectively, while $Z^{q=0}_{PL}$ and $Z^{vdW=0}_{PL}$ are the configuration integrals of the ligand in the bound phase for the *discharging* and *vanishing* step respectively. However, in order to obtain the standard binding free energy ΔG° it is important to apply a standard state correction term(ΔG_{rest}). In the *vanishing* step, the ligand is restrained to the binding site of the protein. This action is performed in order to prevent the non-interacting molecule from drifting away of the protein's cavity. Depending on the type of the restraint, it can be shown that ΔG_{rest} is proportional to the ratio between the volume explored by the ligand inside the binding site, V, that can be computed numerically, versus the reference standard volume, V°, which is typically 1 M (1661 Å^{-3} mol^{-1}). Thus, the standard binding free energy can be eventually computed as:

$$\Delta G^{\circ} = (\Delta G^{q=0}_{host} + \Delta G^{vdW=0}_{host}) - (\Delta G^{q=0}_{solv} + \Delta G^{vdW=0}_{solv}) + \Delta G_{rest}, \tag{1.59}$$

$$\Delta G^{\circ} = -k_B T ln \frac{Z_{PL}}{Z_L Z_P} \frac{V}{V^{\circ}} . \tag{1.60}$$

1.4.2 Markov State Models

1.4.2.1 Introduction

As discussed in the Molecular Dynamics Section 1.3, a major challenge of these simulations is to reach biological relevant timescales. This is due to the extremely small timestep (fs) compared to the timescales where many of the molecular processes of interest typically occur (μs to s). Many techniques have been proposed to solve the timescale problem. One approach that has received interest in recent years is Markov State Models (MSM). MSMs provide a way to describe long-time statistical dynamics as a Markovian jump process on a discrete partition of the configurational space. Because of the Markovianity, the probability of jumping from one state to another depends only on the current state (memoryless property). Thus the creation of an MSM model involves the discretisation of the configurational space into a set of n disjoint, discrete states S_1,...,S_n and a nxn transition probability matrix $\boldsymbol{P}_\tau$ = [$p_{ij}(\tau)$] expressing the conditional probability of finding the system in state *j* at time *t*+τ given that it was in state *i* at time *t*. The transition probability matrix is estimated from the MD simulation trajectories x_t as it is illustrated in the following equation:

$$p_{ij}(\tau) = P(x_{t+\tau} \in S_j | x_t \in S_i) , \tag{1.61}$$

where τ is the lag time for which the transition matrix is constructed. Therefore, the transition probability matrix is characterised by the n states and by the chosen lag time.

The key steps to build an MSM are summarised in the following steps:

1. Choose an appropriate distance metric for discretization and cluster the MD trajectories into microstates.

2. Select an appropriate lag time.

3. Estimate the transition probability matrix.

4. Validate the model.

5. Coarse-grain the model to gain human intuition.

1.4.2.2 State decomposition

The first step to build an MSM is the generation of discrete states (microstates) from MD simulations. Each microstate consists of a group of structures that should inter-convert rapidly with each other with respect to other microstates. For this purpose, it is important to choose a distance metric that could best capture the relevant dynamics of the system under scrutiny. The process of transforming the Cartesian coordinates of each frame of the MD trajectories in a kinetically meaningful manner for instance root-mean-square deviation (RMSD) between atoms is called *featurisation*. Optionally, a dimensionality reduction of the feature space can be performed using Time-lagged independent component analysis (TICA).[55] This method performs linear transformation of the input coordinates described by the distance metric into a set of coordinates sorted by "slowness". Therefore, it provides an efficient way to obtain a lower dimensional space that maintains the long-timescale dynamics and can be discretised with higher resolution and higher statistically accuracy.

Once features that can best describe the underlying dynamics of the system have been chosen a variety of algorithms may be used to cluster the conformations (such as k-centers, k-medoids and k-means).[56] In this research the k-means algorithm was generally used. It can be described by the following protocol[57],[58]:

1. Randomly choose k-conformations k_1,k_2,...,k_K as the initial centers of K microstates S_1,...,S_K.

2. Calculate the Euclidean distance between every conformation ($\boldsymbol{X}_i$) of the dataset to each of the microstate centers

3. Assign each conformation to the microstate with the minimum distance.

4. Update the new centers of the microstates using the mean vectors. Each mean vector is calculated from the average of the distances between all conformations of each cluster.

5. Repeat steps 2,3 and 4 until the following equation for each cluster is minimised:

$$E = \sum_{k=1}^{K} \sum_{\mathbf{X}_i \in C_K} |\mathbf{X}_i - S_K|^2 \tag{1.62}$$

   where E is the sum of the squared distance errors of all microstates.

1.4.2.3 Estimation of the transition probability matrix

Once data has been assigned to clusters, we can count the number of transitions between each pair of states at an appropriate lag time τ and store them as a transition count matrix C, where C_{ij} is the number of observed transitions from state i to state j. With infinite data, a reasonable estimate (maximum likelihood estimate) to convert C into a transition probability matrix, $p_{ij}(\tau)$ is:

$$p_{ij}(\tau) = \frac{C_{ij}}{\sum_j C_{ij}}. \tag{1.63}$$

However, in practice, a number of issues arise with this approach. The first one is the generation of the transition count matrix C. The first step to count the transitions, is to select an appropriate lag time τ. For this purpose, we have to estimate transition probability matrices in different lag times and for this reason we cover the process of estimating these matrices first. In order to count the transitions we use the sliding window approach. In this method, we generate the transition count matrix C using all the available data. We assume that the conformations were sampled at a regular time interval Δt, where $\Delta t < \tau$. Then the transitions between states are counted as $\sigma(0) \rightarrow \sigma(\tau)$, $\sigma(\Delta t) \rightarrow \sigma(\Delta t + \tau)$, $\sigma(2\Delta t) \rightarrow \sigma(2\Delta t + \tau)$..., where $\sigma(t)$ is the state index of the simulation at time t.[59] This approach will give a more accurate estimate of maximum likelihood transition matrices, but will underestimate the uncertainty of the model.

A second possible problem with the standard estimate of p_{ij} is satisfying the detailed balance as expressed in Equation 1.64:

$$\pi_i p_{ij} = \pi_j p_{ji} , \tag{1.64}$$

where π_i is the stationary probability of state *i*. This relationship states that for every transition from state *i* to *j*, there should also be a reverse transition from *j* to *i*. One way to enforce this property is using a maximum likelihood method for the estimation of the best reversible transition matrix for the observed data. In this work, a Bayesian reversible MSM estimation was used as it was described in Trendelkamp-Schroer *et al.*[60]

Finally, the states used for the estimation of the matrix should be fully connected. This means starting from any state, every other state can be reached given enough time. A non-ergodic model can arise from MD simulations with different initial configurations that do not overlap due to insufficient sampling. This problem can be solved with longer time scale MD simulations.

The resulting transition matrix can provide relevant and interesting information about the system through its eigendecomposition. This process results in a set of eigenvectors, ψ_i and their corresponding eigenvalues, λ_i:

$$P_\tau \circ \psi_i = \lambda_i(\tau)\psi_i . \tag{1.65}$$

As P_τ is a reversible matrix all the eigenvalues are real with values ranging from -1 to 1, -1 < $\lambda_i \leq 1$. The highest eigenvalue, λ_i, is equal to 1 and its corresponding eigenfunction, ψ_1, is the stationary distribution, π, which consists of the equilibrium probabilities of the microstates:

$$\pi^\top P_\tau = \pi^\top . \tag{1.66}$$

All the other eigenvalues are related to the relaxation timescales within the system, indicating how quickly the process decays towards equilibrium (positive eigenvalues) or oscillates (negative eigenvalues). The positive eigenvalues can be converted to characteristic or implied timescales of the dynamical processes within the system using the lag time τ as shown by the following equation:

$$t_i = \frac{\tau}{ln|\lambda_i|}, \tag{1.67}$$

where t_i is the relaxation time of the i^{th} process determined by the i^{th} largest eigenvalue λ_i. The corresponding eigenvectors are related to the dynamical processes themselves and their coefficients indicate the structural changes that occur in the system during the process.

1.4.2.4 Validation of the Markov Model and lag time selection

It is important to select a proper lag time τ to estimate the transition probability matrix. This should be long enough to ensure Markovianity in state space but also short enough to resolve the system dynamics. Plotting the implied timescales as a function of the lag time (Equation 1.66) can be used as a diagnostic tool to select an appropriate MSM lag time. This plot should give an indication of the smallest lag time needed to satisfy the Markov assumption. Beyond this time, the implied timescales should be converged as a function of τ and thus independent of lag time τ.[61],[62]

Once the transition probability matrix for a chosen lag time has been computed its Markovianity can be tested using the Chapman-Kolmogorov (CK) equation[59],[63]:

$$P(n\tau) = P(\tau)^n. \tag{1.68}$$

This relationship shows that a transition probability matrix estimated at lag time $n\tau$, where n is an integer greater than 1, should be equivalent to the transition probability matrix, P_τ, to the n^{th} power.

1.4.2.5 Coarse-graining for the generation of macrostates

The MSM, created from the estimation of the transition probability matrix and validated by the CK test, contains hundreds or thousands of microstates that can approximate the statistical dynamics of a biomolecular system. However, in order to obtain an interpretative model it is useful to construct metastable macrostates from a kinetic lumping of the microstates. Metastable states contain a collection of microstates that are kinetically related to each other. They can be identified from a relative large gap in the implied timescales plot. There are several ways to perform the coarse-graining of the states, but in this work a more robust version of Perron Cluster Cluster Analysis (PCCA+) algorithm compared to simple PCCA was chosen.[64–66] In PCCA, we start off with all microstates combined into a single macrostate and then sequentially break this macrostate into two smaller macrostates based on the next slowest right eigenvector.[67] If there is a clear gap in the separation of the timescales for the different metastable states, then the model will be very useful for human intuition. However, in many cases there is a continuum of eigenvalues that can lead to a propagation of error. This error arises from the fact that not all eigenvalues may participate strongly in each eigenmode, therefore many microstates will be assigned to macrostates arbitrarily. PCCA+ tries to tackle this error by simultaneously considering the relevant slowest dynamical eigenvectors.

After coarse-graining microstates to macrostates, it is a good practice to test the quality of the macrostate model by comparing the relaxation times of the model with those of the microstate model and examining how close they are. For every macrostate we can approximate the stationary probabilities from its transition probability matrix. In addition, the Mean First Passage Time (MFTP), which is the average time taken to get from a macrostate i to the macrostate j, can be defined as[68–70]:

$$MFTP_{ij} = \sum_{j=1}^{N} P_{ij}(\tau)(\tau + MFTP_{ji}) \,, \tag{1.69}$$

where N is the number of discrete states. The boundary condition for this calculation is that:

$$MFTP_{jj} = 0 \,. \tag{1.70}$$

Thus, the set of linear equations from 1.69 and 1.70 can be solved to obtain $MFTP_{ij}$.

1.5 Experimental techniques used to investigate protein-ligand binding affinity

1.5.1 General Overview

The binding of small molecules to proteins can be investigated through many experimental techniques. For instance, X-ray crystallography and cryo-electron microscopy can provide atomic resolution or near atomic resolution structures of protein-ligand complexes.[71] In addition, NMR spectroscopy can be used to characterise the dynamics of the binding process over a wide range of timescales from picoseconds to seconds.[72] Finally, the thermodynamic parameters of binding events, i.e. the heat change that occurs when biomolecules interact, can be measured either directly using Isothermal Titration Calorimetry (ITC) or indirectly by using techniques that can calculate the binding affinity as a function of temperature such as Surface Plasma Resonance (SPR) and Fluorescence Polarization (FP).[73] These three techniques will be introduced and discussed in detail in the following subsections.

1.5.2 Isothermal Titration Calorimetry (ITC)

ITC is a quantitative technique that measures the heat change during molecular association at a constant temperature. ITC monitor these heat changes by measuring the power needed to maintain the same temperature between two identical cells, made of a highly efficient thermally conducting material, as protein and ligand are mixed. Usually, the sample cell contains the protein of interest and the reference cell is filled with water or buffer. ITC is considered as the gold standard in characterizing interactions of biomolecules in a broad range of binding affinities, because it provides a full thermodynamic description of the system of interest in a single experiment.[71]

During an ITC experiment, the ligand is titrated into the sample cell in precisely known aliquots. This results to heat release since binding is an exothermic reaction. This causes a temperature imbalance between the reference and the sample cell that is compensated by changing the feedback power provided to the cell heater. The power applied to the sample cell in order to maintain equal temperatures between the two cells at each titration is plotted against time (primary ITC data). These data are then normalised for concentration to produce a titration curve of kcal/mol versus molar ratio of the total ligand concentration to the protein concentration. Finally, this titration curve is fitted to a binding model in order to obtain the binding constant, Gibbs free energy of binding, binding enthalpy and the stoichiometry (n) of the binding event.[74],[24] An example of the ITC data is provided in Figure 1.8.

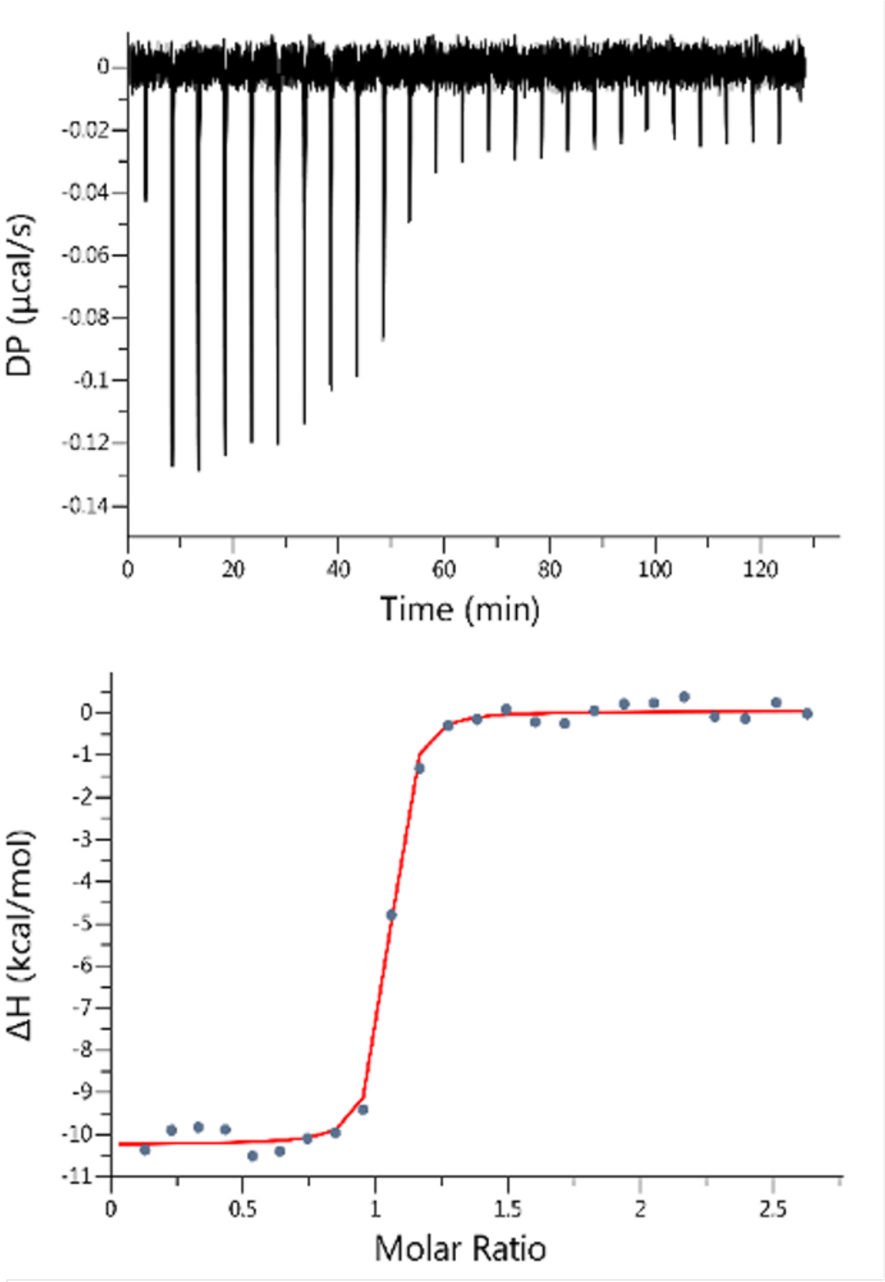


FIGURE 1.8: A representative example for ITC data. Primary ITC data, showing observed changes in heat resulting from interactions between the biomolecules are depicted in the upper panel. The resulting binding curve from the fitted binding model is illustrated in the lower panel.

1.5.3 Surface Plasma Resonance (SPR)

SPR is a biophysical approach for the study of protein–ligand binding kinetics and affinities.[75] It requires protein molecules (receptor) to be immobilised on a sensor surface, which is usually a thin film of gold on a glass support that forms the floor of a small-volume flow cell through which an aqueous solution flows continuously. The examined ligand (analyte) is injected in the aqueous solution through the flow cell in order to measure the binding reaction. If the analyte binds to the receptor there is an association phase during which the binding sites are occupied and an increase in the refractive index at the surface (expressed in response units, RU) is observed. This increase is measured as a function of time resulting a time-dependent RU curve that can be used to calculate the kinetic association rate constant, k_{on}. After a desired period of time, a buffer solution containing no analyte is injected through the flow cell causing the dissociation of the protein–ligand complex and to a decrease in the refractive index. Similarly, a second time-dependent RU curve is produced and can be used to measure the rate of dissociation k_{off}. Finally, the binding constant K_b can be obtained according to Equation 1.3.[76],[24]

1.5.4 Fluorescence Polarization (FP)

FP is a fluorescence-based method used to measure the kinetics and the thermodynamics of protein-ligand binding. The principle of FP derives from the fact that the degree of polarisation of a small molecule is related to its molecular rotation. In addition, the polarization lifetime depends on the rotational relaxation time, i.e, the time that a molecule needs to rotate through an ca. 68.5 angle after excitation. The latter is proportional to the molecular volume of the examined molecule and thus the ligand in solution unpolarizes faster compared to when its bound to a protein. This allows FP to be used to measure the association of a fluorescent ligand with a larger molecule. FP is assumed to be linear proportional to the percentage of bound/free species and this can be used to quantitatively determined the IC_{50} value of protein-ligand binding.[77],[78] Finally, the corresponding $\mathrm{K}_i/\mathrm{K}_d$ ratio, where K_i is the inhibition constant of the unlabelled small molecule, can be calculated using the appropriate versions of the Cheng-Prusoff equation.[79],[80]

1.6 Statement of aims

Molecular Dynamics (MD) simulations have become popular in the context of free energy calculations of protein-ligand complexes to assist drug discovery. However, MD methods struggle to explore potential energy landscapes efficiently, as they can sample only the low energy conformations of the system of interest. The main goal of this thesis is to examine in depth the sampling problem of MD methods in the calculation of standard free energies of binding of different biomolecular complexes.

Chapter 2 will discuss the use of MD simulations as a sampling technique for the implementation of a Relative Free Energy (RelativeFEP) protocol using the Sire/OpenMM Molecular Dynamics (SOMD) software. RelativeFEP can accurately predict the binding selectivity of a congeneric series of ligands to a protein-target. This is due to the reduced sampling needed for these type of calculations, as they consider only the relative free energy differences of two structural similar molecules that adopt the same binding modes. The efficiency of this method was examined as a part of a novel ligand design workflow to optimize the binding selectivity of cyclophilin (Cyp) inhibitors.

Chapter 3 will deal with the validation of an Alchemical Free Energy (AFE) protocol for host-guest binding affinities predictions using the SOMD software. MD simulations are employed again as a sampling technique for the calculation of absolute (standard) free energies of binding. Absolute free energy calculations are a computationally efficient rigorous method to compute binding free energies for a diverse set of compounds.[81] However, they are computationally expensive as they require thorough sampling of the system's degrees of freedom. For this reason, the protocol was initially evaluated in the SAMPL6 challenge in terms of accuracy and correlation with experimental data for a dataset of 27 host-guest systems.

Chapter 4 applies AFE to the challenging case of a ligand binding to the intrinsically disordered protein c-Myc. The AFE results are compared to those obtained with a Markov State Modelling (MSM) protocol.

Finally, Chapter 5 will introduce an adaptive sampling version of the AFE protocol to calculate absolute binding free energies of a diverse set of inhibitors of the flexible protein Murine Double Minute-2 (MDM2) at a reduced computing cost. The protocol will be tested in terms of its ability to reproduce the binding selectivity of these compounds between a full length and lid-truncated variant of MDM2.

Ultimately, Chapter 6 will draw a summary and conclusions from all the results of this thesis.

Chapter 2

Computationally Driven Discovery of Novel Cyclophilin A and D Inhibitors

2.1 Introduction

2.1.1 General information on Cyclophilins

Cyclophilins (Cyps) are a family of proteins that catalyze the inter-conversion of *cis* and *trans* isomers of proline residues. Cyps were initially identified in 1984 when Fischer *et al* found a protein, which they called peptidyl-prolyl cis-trans isomerase (PPIase), that increased the rate of the 180° rotation about the (C-N) linkage of the peptide bond of proline.[82] Later, in 1989 it was shown that this protein was the same as the already identified protein CypA. Together with the FK509 binding proteins, they were initially described as the biological receptors for the drugs Cyclosporin A (CsA), FK506/tacrolimus and rapamycin/sirolimus (Figure 2.1).[83]

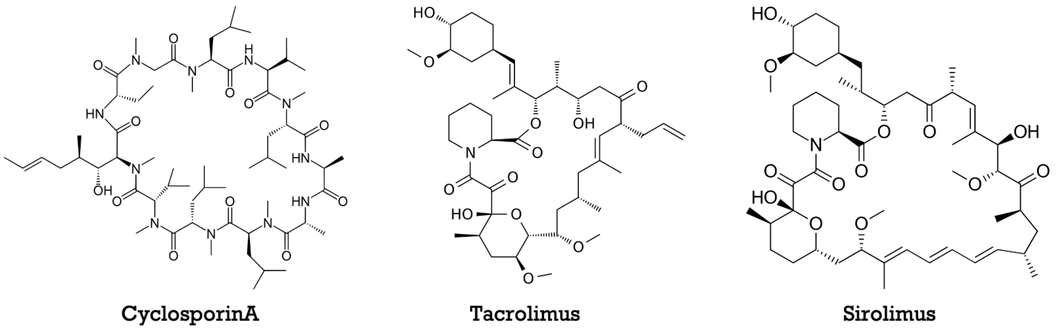


FIGURE 2.1: Chemical structures of CsA, Tacrolimus and Sirolimus.

The peptidyl-propyl isomerisation is essential for many processes such as protein folding and assembly of multidomains. The peptide-bond has a partial double bond character and can exist in two different forms: *cis* and *trans*. For non-prolyl bonds, the *trans* conformation, where the side chains are 180° opposite to each other, is sterically favored and the equilibrium constant [trans]/[cis] (K_c) is usually higher than 100 ($K_c > 100$).[84],[85] However, due to proline's unusual cyclic structure, both the *cis*/*trans* conformations of the peptidyl-propyl bond are accessible and the *trans* isomer is only slightly favored. It has been demonstrated in refolding experiments, that the peptidyl-propyl bond will not adopt the intended conformation spontaneously due to the relatively high activation energy required to catalyse the *cis*/*trans* isomerisation (ca. 20 kcal mol^{-1}). Thus, the *trans* to *cis* isomerisation is a slow process and can be the time-limiting step of folding especially in low temperatures.[86] Cyps are enzymes that stabilize the *cis*/*trans* transition state and accelerate the isomerisation (Fig 2.2).[87]

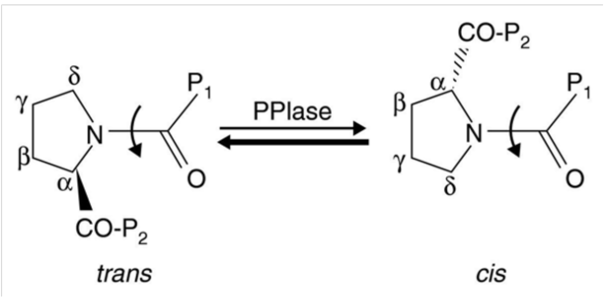


FIGURE 2.2: Schematic representation of the *trans* and *cis* isomers of the peptide bond between proline (illustrated on the left of each structure) and another amino acid (shown as P1, on the right). P2 is a third amino acid bound on the other side of the proline. The carbon atoms of the proline are labelled using Greek letters. The peptide bond is planar and has a partial double bond character. The catalysis of the time-limiting *trans* to *cis* isomerisation of the peptide bond is accelerated by cyclophilins and other PPIases. Adapted from Wang *et al.*[87]

The human Cyp family consists of 17 members with CypA being the most abundantly expressed. They can also be found in mammals, plants and parasites. The majority of these proteins have unknown function and only 7 of them to date have been demonstrated to bind to CsA and/or to posess an isomerase activity, namely CypA, CypB, CypC, CypD, CypE, Cyp40 and CypNK. The aforementioned Cyps differ in their structural characteristics and their localization in the human body. However, they share a common PPIase domain of approximately 109 amino acids, which is surrounded by unique domains for each member of the family that are vital for their sub-cellular compartmentalization and functional specialization. All Cyps have the same secondary and tertiary structures [88] consisting of eight antiparallel β-sheets with two α-helices that pack the β-strands. The active binding site of Cyps is formed by the catalytic amino acid Arg55 and a mixture of hydrophobic, aromatic and polar residues including Phe60, Met61, Gln63, Ala101, Phe113, Trp121, Leu122 and His126. The chemical structure of CsA and a 3D representation of CypA illustrating the key amino acids of the binding site determined by CsA is shown in Figure 2.3:

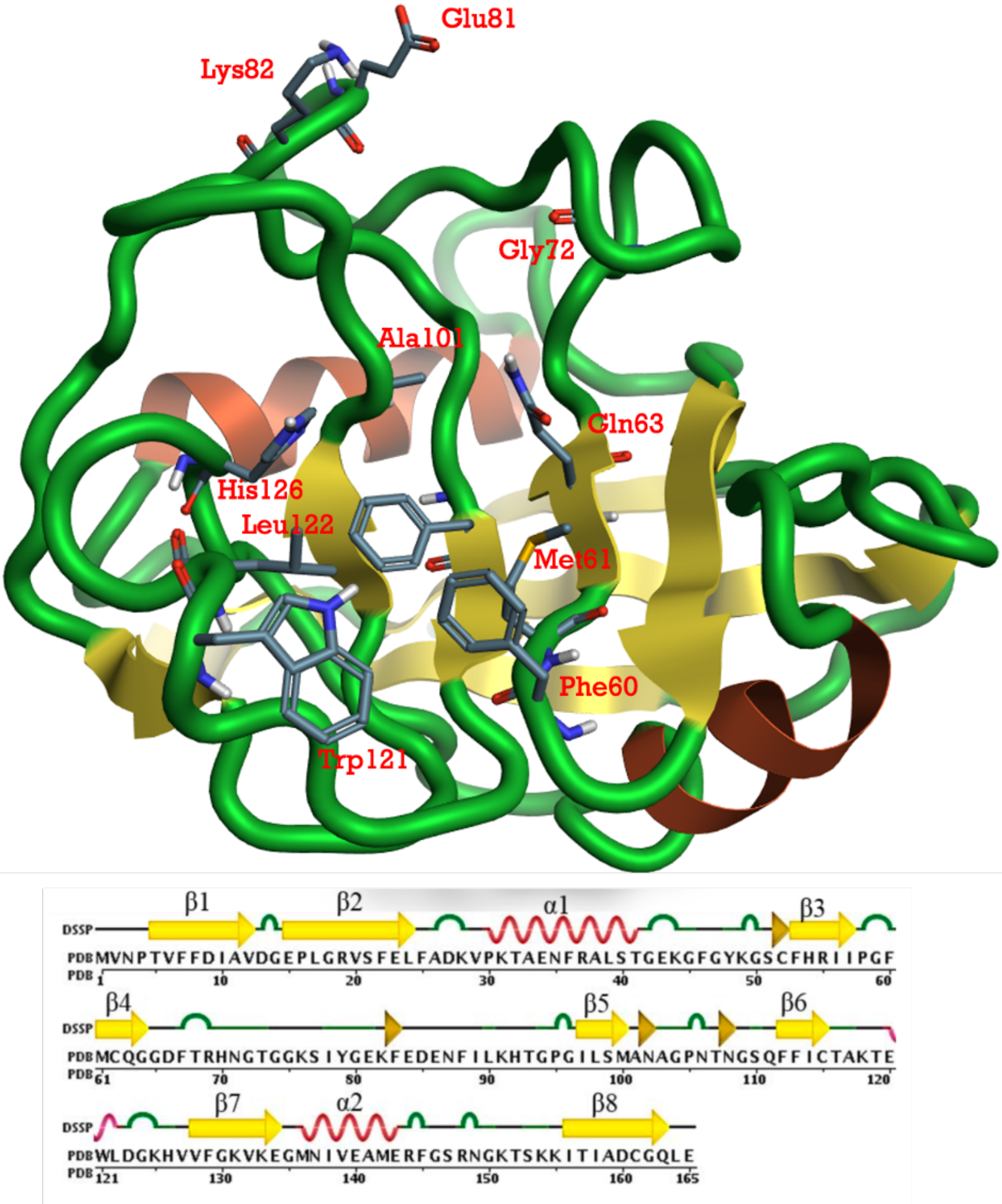

Glu81
Lys82
Gly72
Ala101
Gln63
His126
Leu122
Met61
Phe60
Trp121
β1
β2
α1
β3
DSSP
PDB MVNPTVFFDIAVDGEPLGRVSFELFADKVPKTAENFRALSTGEKGFGYKGSCFHRIIPGF
PDB 1 10 20 30 40 50 60
β4
β5
β6
DSSP
PDB MCQGGDFTRHNGTGGKSIYGEKFEDENFILKHTGPGILSMANAGPNTNGSQFFICTAKTE
PDB 61 70 80 90 100 110 120
β7
α2
β8
DSSP
PDB WLDGKHVVFGKVKEGMNIVEAMERFGSRNGKTSKKITIADCGQLE
PDB 121 130 140 150 160 165

FIGURE 2.3: 3D representation and amino acid sequence of CypA. The key residues forming the active binding site of CypA are colored red. Alignment of sequence and secondary structure was obtained from http://www.rcsb.org/.

Many residues are highly conserved amongst the Cyp family as is illustrated in Table 2.1. [87]

Protein name	Identity(%)	**Binding site residues of Cyp PPIase domain**												
(other name)	to PPIA	55	60	61	63	72	81	82	101	102	110	121	122	126
PPIA (CypA)	-	Arg	Phe	Met	Gln	Gly	Glu	Lys	Ala	Asn	Ser	Trp	Leu	His
PPIB (CypB)	64	Arg	Phe	Met	Gln	Gly	Glu	**Arg**	Ala	Asn	Ser	Trp	Leu	His
PPIC (CypC)	63	Arg	Phe	Met	Gln	Gly	Glu	**Thr**	Ala	Asn	Ser	Trp	Leu	His
PPID (Cyp40)	60	Arg	Phe	Met	Gln	Gly	Glu	Lys	Ala	Asn	Ser	**His**	Leu	His
PPIE (CypE)	67	Arg	Phe	Met	Gln	Gly	**Lys**	Lys	Ala	Asn	Ser	Trp	Leu	His
PPIF (CypD)	76	Arg	Phe	Met	Gln	Gly	**Ser**	**Arg**	Ala	Asn	Ser	Trp	Leu	His
PPIG	52	Arg	Phe	Met	Gln	Gly	**Gly**	**Phe**	Ala	Asn	Ser	**His**	Leu	His
PPIH	53	Arg	Phe	Met	Gln	Gly	**Gly**	**Pro**	Ala	Asn	**Cys**	Trp	Leu	His
PPIL1	54	Arg	Phe	Met	Gln	Gly	**Lys**	**Gln**	Ala	Asn	Ser	Trp	Leu	His
PPIL2	49	Arg	Phe	**Val**	Gln	Gly	**Lys**	**Pro**	Ala	Asn	Ser	Trp	Leu	His
PPIL3	50	Arg	Phe	Met	Gln	Gly	**Lys**	Lys	Ala	Asn	Ser	**His**	Leu	**Tyr**
PPIL4	36	Arg	Phe	**Ile**	Gln	Gly	**Gly**	**Leu**	**Val**	Asn	Ser	**Tyr**	Leu	His
PPIL6	43	Arg	**Gly**	Met	Gln	Gly	**Pro**	**Thr**	Ala	Asn	Ser	**Tyr**	Leu	**Phe**
PPWD1	49	Arg	Phe	Met	Gln	Gly	**Gly**	**Glu**	Ala	Asn	Ser	Trp	Leu	His
NKTR	50	Arg	Phe	Met	Gln	Gly	**Gly**	**Tyr**	Ala	Asn	Ser	Trp	Leu	His
CWC27	43	Arg	Phe	**Ile**	Gln	Gly	**Ala**	**Pro**	Ala	Asn	Ser	**Glu**	Leu	His
RANBP2	66	Arg	Phe	**Val**	Gln	Gly	**Asp**	Lys	Ala	Asn	Ser	Trp	Leu	His

TABLE 2.1: Percentage sequence identity of all the Cyclophilins' PPIase domain compared to CypA. Comparison of the conservation of the amino acids that characterise the binding site that was defined by CsA. The residues of the Cyclophilin members that differ from the ones of CypA are illustrated in bold. The data are provided from Charis Georgiou.

2.1.2 Biological role of Cyps

Cyps are the intracellular receptors for CsA, a cyclic 11-amino acid peptide originally isolated from the fungus *Tolypocladium inflatum*. CsA is an immunosuppressive drug that is used in organ-transplant patients to prevent immune response and organ rejection. The major *in vivo* receptor is CypA and the resulting CsA-CypA complex has the ability to bind and inhibit calcineurin. As a consequence pNFAT, the calcineurin substrate, is unable to translocate from the cytosol to the nucleus and activate the T-cells.[89] Moreover, Cyps are responsible for the misregulation of diverse biological signaling pathways such as RNA splicing[90] and mitochondrial apoptosis. They are also involved in the life cycle of different viruses for instance Human Immunodeficiency Virus (HIV-1) and Hepatitis C Virus (HCV) and in different types of cancer.[91],[92]

The present work focused on the biological role of two members of the Cyp family, CypA and CypD. CypA has a molecular mass of 18 kDa and is one of the most abundant proteins in the cytoplasm and is involved in a plethora of cellular functions such as protein folding, trafficking and cell signaling.[93] CypA is not necessary for cell growth and survival[94],[92], but its extracellular fraction acts as pro-inflammatory mediator that triggers inflammatory responses and probes chemotactic activity for neutrophils and monocytes via the CD147 receptor.[95]

CypA is reported to be overexpressed in numerous types of cancer. Previous studies have emphasized that CypA can be identified as a bio-marker in lung cancer[96] and that overexpression of CypA in lung cancer cells increases cancel cell growth, whereas knockdown of CypA slows down the cell growth.[97] Furthermore, CypA interacts with CD147 and stimulates the human pancreatic cancer cell proliferation.[98] Regarding breast cancer, CypA regulates the Jak2/Stat5 pathway which is vital for the tumorigenesis. In addition, CypA has been spotted as a useful Hepatocellular carcinoma marker[99] and has been overexpressed in primary and metastatic melanoma[100] and in glioblastoma multiform[101]. Qi *et al* also suggests that CypA can have a possible role in malignant transformation of esophageal squamous cells[102] and Choi *et al* showed that upregulation of

CypA in prostate cancer cells provides resistance to cisplatin and hypoxia-induced cell death.[103] Finally, CypA is associated with tumor progression and tumor development in colocteral cancer.[104] To summarize, CypA plays a vital role in tumor development and is overexpressed in numerous types of cancer.[105]

CypA is also involved in HIV and HCV infections. Regarding the HIV-1 infection, CypA can interact with HIV capsid proteins on their CA domain, especially with a proline containing sequence in the capsid polyprotein Gag[106] as well as with HIV accessory proteins such as the viral protein R (Vpr).[107],[108] Concerning on HCV virus, CypA interacts with non-structural protein 5A (NS5A) and the binding site has been located to the proline rich domain II, centered around a "DY" dipeptide motif that controls CypA dependence and CsA resistance.[109],[110],[111] This protein is important for HCV replication, because it maintains the proper structure and function of HCV replicase.[112] The CypA-NS5A interaction is conserved among all HCV genotypes and all the cyclophilin inhibitors prevent the formation of this complex.[113]

CypA has also been reported to be involved in many other diseases including cardiovascular diseases, diabetes, other viral and protozoan infections, amyotrophic lateral sclerosis (ALS), rheumatoid arthritis (RA), sepsis, asthma and periodontitis.[93]

CypD (PPIF) is also a member of the Cyps family with important biological role. CypD is located in the mitochondrial matrix and it regulates the opening of the mitochondrial permeability transition pore (mPTP) in response to various stress stimuli.[114] Mitochondria control the Ca^{2+} concentration in different cell sections. mPTP has a vital role for the Ca^{2+} efflux from mitochondria to the cytosol across the inner membrane.[115] This pore in the normal state is very important for cell metabolism, but persistent opening induces necrotic cell death. In 2005, Linkerman *et al* provided evidence that CypD has a key role for necrotic signalling by deleting Ppif, the CypD gene. The cells without this gene were resistant to cell death caused by overload of cytosolic calcium.[116] Figure 2.4 illustrates the mechanism of action of CypD in the mPTP target.

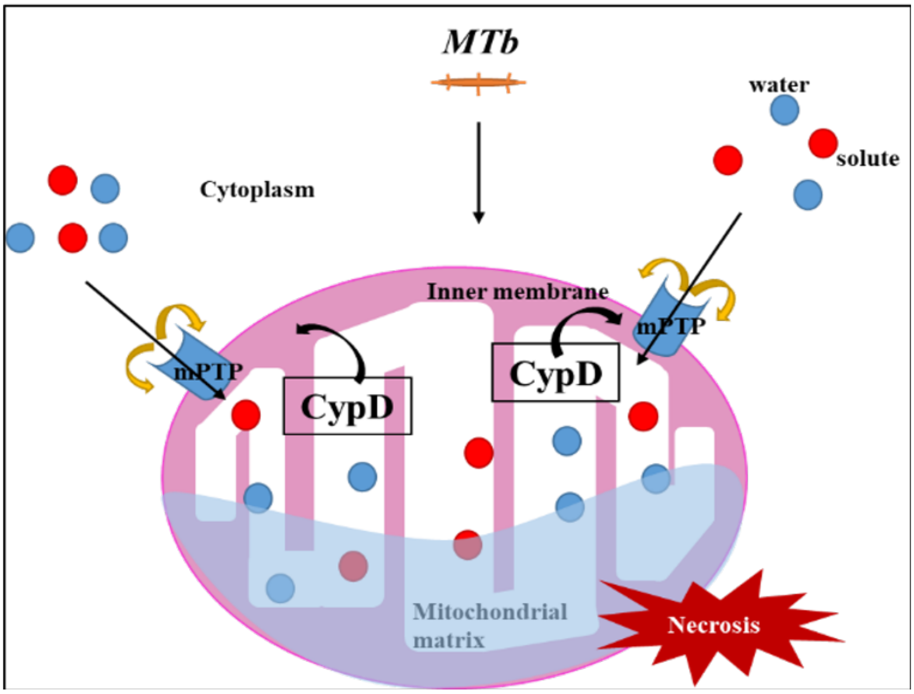


FIGURE 2.4: Schematic illustration of the pharmacological role of CypD in the mPTP opening. CypD translocates from the mitochondrial matrix to the inner mitochondrial membrane where it triggers the opening of the mPTP. This causes the entrance of water and solute (blue and red circles) within the mitochondrion, which then leads to necrosis. (Provided by Maria Kouridaki)

CypD has been considered as a potential biological target for diverse diseases where mitochondrial dysfunction plays a vital role to their pathogenesis. Mitochondrial dysfunction is associated with a variety of liver diseases, as mitochondria play a vital role in the integrity and the normal function of liver cells. Small for size liver syndrome (SFSS) is a clinical syndrome that follows liver transplantation and hepatectomy.[117] Small for size livers are associated with liver cell necrosis and an increase in alanine transaminase and bilirubin levels. A CypD inhibitor (NIM81) decreased graft injury and increased liver regeneration in an experimental model. In addition, it reducted lung inflammation by reducing the expression of inflammatory cytokines and adhesion molecules (tumor necrosis factor (TNF-a) and intercellular adhesion Molecule-1 (ICAM-1)).[118]

Moreover, mitochondrial dysfunction is considered as the triggering event in the development and evolution of nonalcoholic steatohepatitis (NASH) and Non-alcoholic fatty liver disease (NAFLD).[119],[120] NAFLD is one of the most common chronic liver diseases in the world and NASH is the second most common indication for liver transplantation in the USA after chronic hepatitis C.[121] NASH currently affects 3-4% of the US population and by 2020 will be the leading cause of liver transplantation.[122] There are currently no drugs for this disease and it is tackled through combination therapies across a broad range of patients.[123] Wang *et al* provided evidence that overexpressed CypD leads to hepatic steatosis.[124] The proposed mechanism is that persistent mPTP opening and Ca^{2+} balance disruption results in endoplasmic reticulum (ER) stress through p38 mitogen-activated protein kinase (MAPK) activation. Hence, this causes an increased sterol regulatory element-binding Protein-1c (SREBP-1C) and eventually steatosis in the liver.[124] Therefore, pharmacological inhibition of mPTP has been shown to be beneficial in *in vitro* and *in vivo* disease models of liver fibrosis.[113]

Furthermore, mitochondrial dysfunction is connected also to acute pancreatitis (AP). AP is one of the most common pancreatic diseases and it is caused by gallstones or redundant alcohol intake. Severe AP is characterised by pancreatic necrosis, systematic inflammatory response syndrome, multiple organ failure and sepsis which results in the death of 25% of patients.[125],[126] mPTP opening is central to numerous forms of AP and causes reduced adenosine 5′-triPhosphate (ATP) production, defective autophagy, zymogen activation, cytokine release and necrosis.[127] Studies on CypD knockout mice showed that inhibition of CypD can reduce or ameliorate local and systematic pathological responses of AP.[128],[129]

Ultimately, mitochondrial dysfunction is also linked to Alzheimer's disease. Du *et al* showed that CypD interacts directly with Aβ, which plays a vital role in Alzheimer's pathogenesis, in the mitochondria of Alzheimer's disease brain and in a mouse model of Alzheimer's disease. This interaction promotes reactive oxygen species (ROS) generation and recruitment of CypD in the mitochondrial inner membrane. This results in persistent opening of mPTP that regulates the mitochondrial-induced cell death in an Aβ-rich environment. Inhibition of CypD protected neurons from Aβ- and oxidative stress-induced cell death, improved learning and memory and synaptic dysfunction.[130]

2.1.3 Current Cyp inhibitors and the need for the discovery of new ones

Due to their diverse biological role, as mentioned above, Cyps are considered as potential therapeutic targets for tackling numerous diseases. The story of Cyp inhibitors began in 1969 in Norway when CsA and Cyclosporin C (CsC) were isolated from the fungus *Tolypocladium inflatum*. In 1983, it was developed by Novartis as a drug in organ transplantation.[131] However, the long-term use of CsA in organ-transplant patients caused severe side effects such as nephrotoxicity and this is the major obstacle for the broader use of this drug.[132] This study led many researchers to find CsA analogues that lack calcineurin-binding properties and therefore do not exhibit immunosuppressive properties. Three major semi-synthetic analogues of CsA were proposed and tested in clinical trials for the treatment of viral infections namely Alisporivir (DEB025), NIM811 and SCY-635 (Figure 2.5).

Alisporivir is a semi-synthetic analogue of CsA with increased inhibitory activity over CypA,[133] and the ability to inhibit the calcineurin binding.[134] It was initially developed as a drug for HIV infection, but it has entered a phase II trial for patients infected with genotype 2 or genotype 3 HCV as a monotherapy or in combination with interferon and ribavirin.[92],[135],[136] NIM811 has higher affinity for CypA than CsA and has been investigated clinically as a potential treatment for HCV.[137] Combined with NS3-4A protease or NS5B polymerase inibitors it has additive inhibition to viral replication and a high genetic barrier to viral resistance development.[138] In addition, it has anti-HIV activity as it inhibits the binding of CypA to HIV-p24gag protein.[139] Ultimately, it can be produced on large scale directly from fermentation through a genetic manipulation of the producing strain of *Tolypocladium inflatum*.[136] SCY-635 was discovered at Aventis for the treatment of HIV, but it has progressed phase II development for HCV infection.[140] It is slightly more potent than CsA for CypA inhibition and it suppress the HCV replication in replicon cells in a time dependent manner.[141]

In addition to CsA-derived inhibitors, the sangliferins represent another class of macrocyclic natural products that inhibit Cyps. Sangliferins are produced by soil *Streptomyces* bacteria.[142] They have immunosuppresive activity with a mechanism of action that does not involve calcineurin binding, with details yet to be determined.[143],[144] The most abundant member of this class of molecules, Sangliferin A, binds to CypA 60-fold more potently than CsA and can suppresses HCV replication.[145],[146] These data encouraged different groups such as Novartis[142] and Bioteca[147] to develop nonimmunosuppresive sangliferin analogues. Ultimately, Gilead Sciences discovered a family of cyclophilin-binding macrocycles that contain the functionality characteristics of the piperazic acid of sangliferins (The piperazic acid occupies the same hydrophobic pocket as 11-Val of CsA).[148] Chemical structures of CsA-derived inhibitors and sangliferinA are illustrated in Figure 2.5.

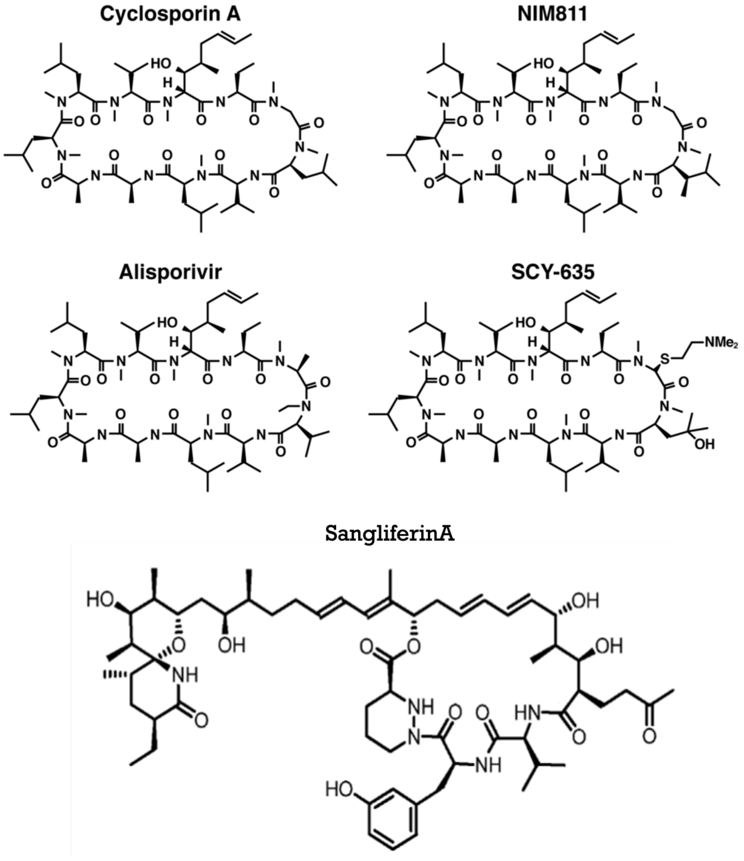
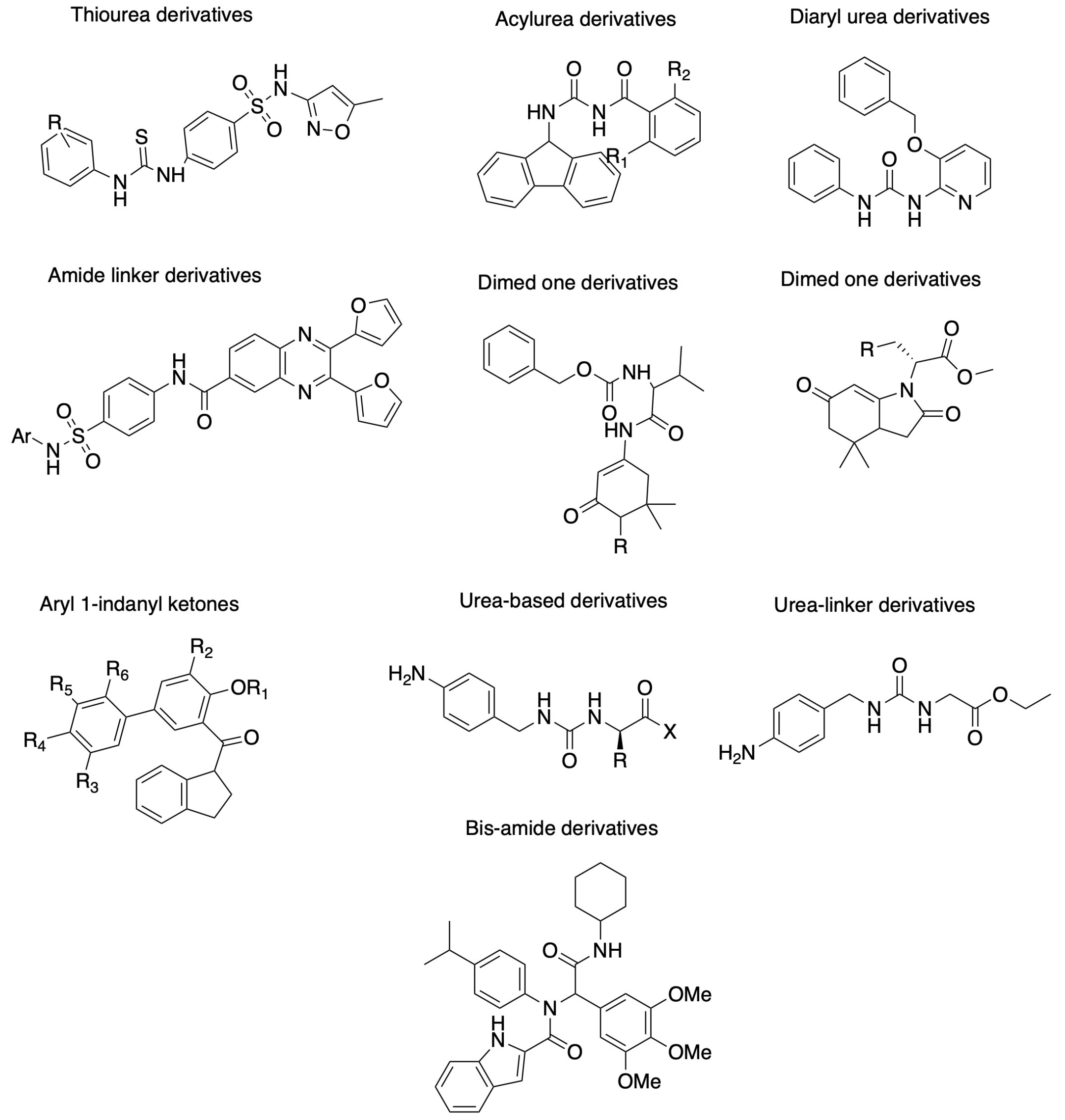


FIGURE 2.5: Depiction of the CsA-derived inhibitors from left to right and top to bottom, Cyclosporin A, NIM811, Alisporivir, SCY-635. Sangliferin A is illustrated at the bottom of the figure. Adapted from Hopkins *et al*.[149]

The main drawback of the aforementioned inhibitors is that they have unfavorable drug-like properties. They are complex to synthesize, lack subtype selectivity and have high molecular weights, limited solubility and poor Central Neural System (CNS) activity.[150] Thus, much synthetic effort has been consumed to develop small molecule inhibitors with improved pharmacokinetic/pharmacodynamic properties (PK/PD). The current strategy for small molecule inhibitors typically includes urea moiety as the central core of the molecule and several urea analogues such as acetyl urea or thiourea.[129],[151],[152] Moreover, Cho *et al* reported amide scaffolds as potentially small molecule inhibitors of Cyps.[153] Finally, Ahmed-Belcasem *et al* identified two compounds with significant inhibition of CypA, CypB and CypD and with anti-HCV activities through a fragment based drug design method by using X-ray crystallography and Nuclear Magnetic Resonance (NMR).[154] Chemical structures of the reported small molecule inhibitors are shown in Figure 2.6.


Thiourea derivatives

Acylurea derivatives

Diaryl urea derivatives

Amide linker derivatives

Dimed one derivatives

Dimed one derivatives

Aryl 1-indanyl ketones

Urea-based derivatives

Urea-linker derivatives

Bis-amide derivatives

FIGURE 2.6: Schematic representation of the small-molecule Cyp inhibitors reported in the literature.

Despite the large number of inhibitors reported in the literature, many of them do not bind with 1:1 stoichiometry and do not have high affinity for

Cyps.[154],[155] However, their main drawback is that they do not offer specific inhibition of Cyp isoforms. Specificity in this family of proteins is very important for the avoidance of side effects from the use of non-specific drug molecules and for the better understanding of the biological role of every isoform. However, obtaining high binding specificity is the most challenging issue in the quest for novel Cyp inhibitors, because of the high degree of similarity of the active site between different Cyp isoforms, as decribed above. The similarity is illustrated with a small inhibitor, compound **1**,[154] in Figure 2.7.

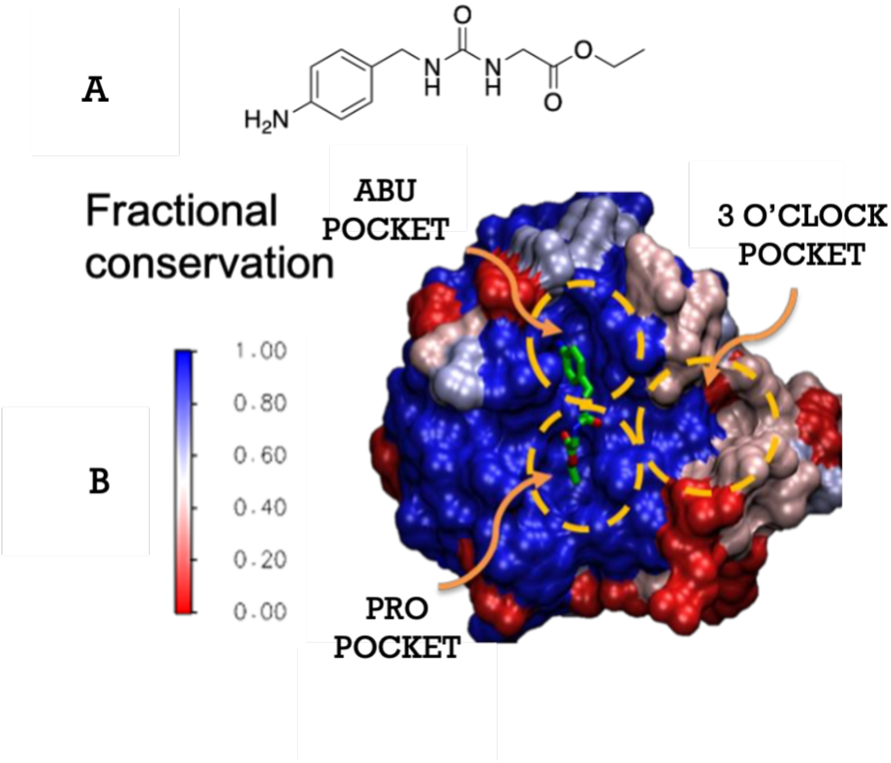


FIGURE 2.7: **A)** Chemical structure of compound **1**.[154] **B)** Three dimensional surface structure of the PPIAse domain of the human CypA-D isoenzymes which is colored by residue conservation (blue highly conserved, red poorly conserved). The location of *Abu*, *Pro* and *3 o'clock* pocket are circled and the small inhibitor is depicted in colored sticks. Adapted from Simone *et al.*[156]

2.1.4 Previous work in the Michel group

Figure 2.7 also highlights a less conserved *3 o'clock* pocket that could be potentially used to achieve selective inhibition amongst different Cyp isoforms. This accessory pocket is located close to the enzyme active site, delineated by the so called *Abu* and *Pro* pockets (these pockets were named after the amino acids *Abu* and *Pro* of CsA that bind to the CypA binding site). Figure 2.8 illustrates the structural difference of the *3 o'clock* pocket in CypA and CypD. In order to target this pocket, De Simone *et al.* examined the binding of the small inhibitor illustrated in Figure 2.7A.[156]

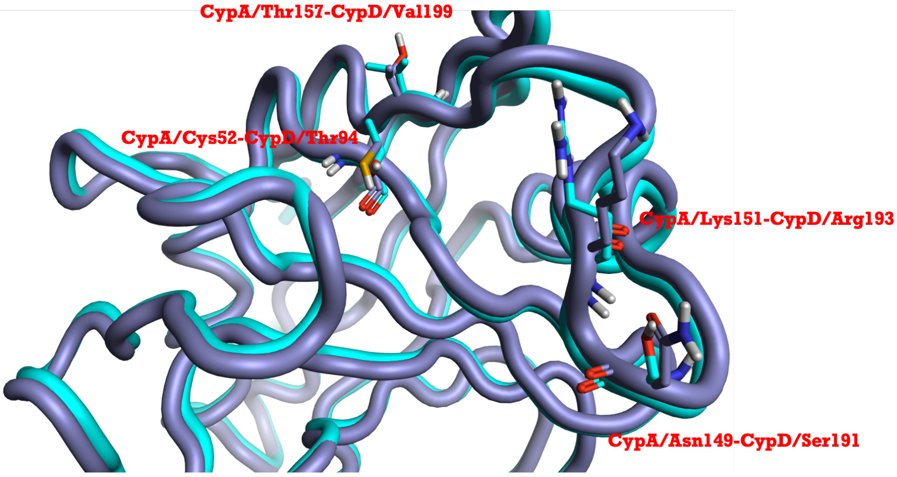


FIGURE 2.8: 3D representation of the less conserved *3 o'clock* pocket in CypA (magenta) and CypD (light blue). The name of the different residues in CypA and CypD is written in red.

The crystal structure of compound **1** in complex with CypA showed that both urea nitrogen atoms have formed hydrogen bonds with the backbone oxygen of Asn102.[154] However, extensive MD simulations on cyclophilins in complex with compound **1** revealed that the nitrogen atom that is further away from the ester moiety interacts only weakly with Asn102 (Fig B.1). Thus, an alkylation of this nitrogen can be tolerated in order to introduce a new vector in the scaffold. In addition, *ab-initio* calculations on model ureas suggest only a small preference to the Z,Z urea conformer over an E,Z conformation. These observations inspired the Michel group to examine if a suitable chosen R group can provide an alkylated urea that would be stable as an E,Z conformer and enable access to the *3 o′clock* pocket (type-II binding mode). (Figure B.2)

The sampling of the desired type-I to type-II binding mode flip is difficult through MD simulations, because the rotational barrier that separates the two conformers is approximately 15 kJ/mol. Thus, the selection of potential R groups to alkylate the nitrogen of the urea was implemented by relative FEP calculations. The protocol used to compute the energetics of this binding mode flip included a perturbation network where the ligands were connected in both binding modes through multiple transformations. The strength of binding to CypA was also examined by Isothermal Titration Calorimetry (ITC).

Compound **1** was predicted to bind with a type-I preference by 1 kcal mol^{-1} based on the FEP calculations reported in De Simone *et al.*[156] Substitutions in the urea nitrogen with non-polar alkyl groups were predicted to favor type-I binding mode and they were less favorable than compound **1**. These results were in line with the experimental ones. Moreover, nitrogen rich five-membered rings (triazole and tetrazole derivatives) illustrated a slight preference in binding mode II based on the computational protocol and more favorable binding than compound **1** in both FEP and ITC. In addition, CypA crystals were soaked with the aforementioned compounds and seven X-ray structures were determined. Based on these results, FEP calculations were proven valid to capture the binding mode preferences of these compounds. Furthermore, the type-II binding mode in the triazole and tetrazole derivatives is suggested to be stabilised by hydrogen bonds that are formed between the nitrogen atoms in the five-membered rings and His54. In addition, this binding mode is supported by a shift of the orientation of the urea carbonyl in order to preserve the placement of the aniline inside the *Abu* pocket. Ultimately, the methyl group of the triazole and tetrazole derivatives appear to project inside the *3 o′clock* pocket.

The projection of the methyl groups of the five-membered rings within the *3 o′clock* pocket and the validation of the binding mode flip hypothesis led to the generalisation of the tri-vector design to other ligand families and Cyp isoforms. For this purpose, a new series of FEP calculations were performed and suggested that replacement of the ester moiety with a bromoaryl pyrrolidine group increased the binding affinity and maintained the preference for a type-II binding mode. Thus, a new family of compounds were generated and assayed by Surface Plasma Resonance (SPR) against three Cyp isoforms (CypA, CypB and CypD). They showed low micromolar to mid nanomolar binding constants and a small degree of isoform selectivity.

Finally, biological assays were performed to confirm the efficacy of the tri-vector design for the new family of ligands. For that purpose, cellular assays were carried out with a triple negative MDA-MB-231 breast-cancer cell line. CypA has been previously reported to be vital for the prolactin-induced activation of Janus-activated kinase 2 in human breast cancer cells [157] and CsA inhibits the growth of the aforementioned cell line. The experimental results illustrated that the new compounds inhibited cell growth in a dose-dependent manner at low micromolar concentration, resulting in potency comparable to CsA. Furthermore, little cell death was observed for these molecules indicating that the inhibition of the human breast cancer cells is due to reduced proliferation rather than as a consequence of cell death as observed in the case of CsA. Additionally, they did not provide evidence of growth inhibition or cell death in the non-tumorigenic fibroblast IMR-90 cell line. This is in contrast to compelling growth inhibition and cell death of CsA in the same cell line. Thus, the new family of compounds were shown to be cytostatic with similar potency to CsA, which is cytotoxic. More information for this study and the chemical structures of the previous ligands can be found in De Simone *et al* paper.[156]

2.1.5 Michel's group lead compound and scope for further improvements

In summary, CypA and CypD have a very important biological role as described in Section 2.1.2. In addition, they are associated with different diseases and this is the reason why selectivity is an essential parameter for small molecule inhibitors. De Simone *et al.* have discovered a novel family of cyclophilin ligands with a unique binding mode that enables targeting of a *3 o'clock* pocket in addition to the usual *Abu* and *Pro* pockets.[156] Compound **15**, illustrated in Figure 2.9, is particularly interesting as it shows reasonable selectivity for CypD over CypA (ca. 8 fold according to SPR assay), and inhibits the growth of MDA-MB-231 cells with GI_{50} values that are two-fold better than CsA. In addition, it offers advantages over the known Cyp inhibitors in terms of ease of synthesis and reduced toxicity.

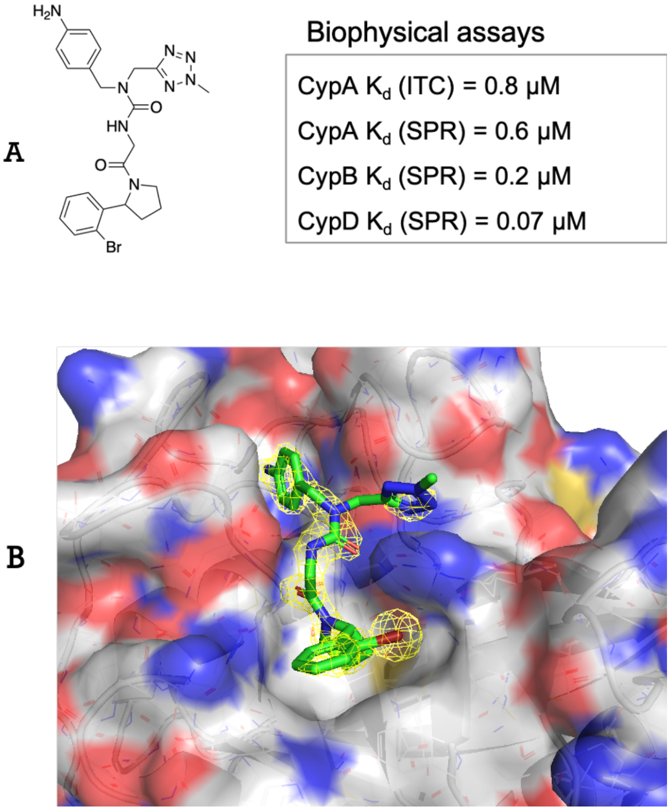


FIGURE 2.9: **A)** Chemical structure of compound **15**. SPR and ITC results for the different Cyp isoforms that suggest a 8-fold preference for CypD over CypA **B)** X-ray crystal structure of CypA in complex with compound **15**. The electron density of the crystal structure is depicted in yellow color.

However, there is scope for further improvements for compound **15**. The aniline ring of compound **15** can lead to the formation of toxic metabolites and thus cause side effects. In addition, the bromine group reduces a lot the solubility of the molecule. Ultimately, there is potential for further derivatisation to extend substituents into the *3 o'clock* pocket and increase the selectivity for one isoform over another. Therefore, the goal of this project is to further improve the design of novel tri-vector Cyp inhibitors and tackle the challenges described above through the following structural modifications depicted in Figure 2.10.

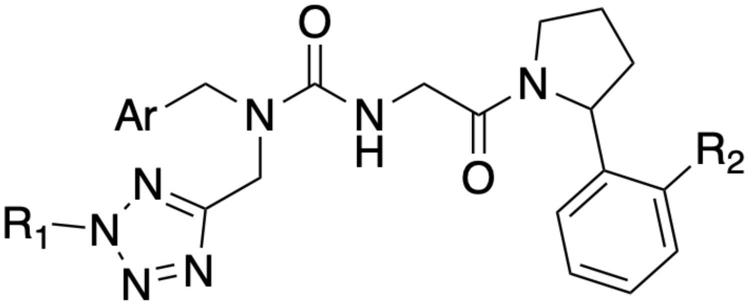


FIGURE 2.10: Envisioned structural modifications on compound **15**.

To achieve this the following targets in the three pockets were identified:

1. Replacement of the aniline ring with another aryl group to block potential oxidation sites that could lead to the formation of toxic iminoquinones. These modifications are going to further reduce the toxicity of compound **15**.

2. Replacement of the bromine with a suitable R_2 group in order to retain the potency and improve the solubility of our lead compound.

3. Replacement of the –Me group by a larger R_1 group should allow extension deeper into the *3 o'clock* pocket for additional gains in potency and selectivity.

We hope to develop two compound based on compound **15**, one with strong dissociation constant for CypD (K_d <10 nM) and >100 fold selectivity for CypD over CypA and one with a K_d lower than 10 nM for CypA bearing at least a 100-fold specificity for CypA over CypD.

2.2 Computational workflow

2.2.1 Construction of Virtual Libraries of Ligands for CypD

To explore the chemical space for binding to the *3 o'clock* pocket, a library of ca. 10.000 analogues of our current lead molecule was constructed. The software Spark from the Cresset company was used for this purpose. Spark uses databases of fragments to suggest replacements in selected regions of a known active molecule.[158] These suggestions aim to preserve the shape and electrostatic properties of these regions. For this purpose, Spark uses the following databases:

1. commercially available compounds from the ZINC library[159],
2. bioactive compounds from literature reports through the ChEMBL database[160] and
3. small molecule X-ray structures from the Cambridge Structural Database (CSD)[161]

For the choice of the appropriate fragments from these libraries a scoring function is needed to evaluate each structure. Spark's scoring function is based on Cresset's field technology [162], that summarises the molecular fields, computed from the eXtended electron distribution (XED) force-field reported in Slater *et al* 2013[163], to the local extrema of the electrostatic, van der Waals (vdW) and hydrophobic potentials of a molecule called *field points*. These points are placed around the known molecule and the resulted molecular structures. Spark's scoring function also uses a shape similarity calculation by Grant *et al*.[164] This algorithm is based on a Gaussian description of molecular shape to compare two molecules. The comparison is performed by using an optimization procedure to maximize the intersection volumes of the examined molecules. Thus, the scoring function takes into account the average of shape and field points similarity of each molecular structure.

A major advantage of Spark is that it scores each potential fragment after it is merged into the starting molecule. Therefore, the procedure for the selection of the appropriate replacements consists of three steps: the choice of fragments with the required number of attachment points and the required shape, energy minimisation of the resulted molecule using the XED force-field to remove steric clashes and unfavorable conformations, and finally application of the scoring function to the whole molecule using an average of shape and field similarity.

Spark offers different ways to perform this workflow. The so called fragment-growing protocol was used for this study, that employs multiple reference compounds: a *Starter* and a *Reference*. The *Starter* molecule is the one to which a modification will be made. In our case, the starting molecule was compound **15** and we wanted to replace the –Me group of tetrazole with larger fragments towards the *3 o'clock* pocket. The *Reference* compounds, provide information about the volume of space that we want to explore and the electrostatic and vdW interactions that the resulted molecules will have. In this study, we wanted to manually construct *Reference* structures that will help us to better explore the *3 o'clock* pocket. We targeted Thr94 of CypD, that extents deeply inside the accessory pocket and offers selectivity over CypA (it contains Cys52 in the same position). For this purpose, two *Reference* compounds were used one with linear substituents and one with a 6-membered ring, and they are depicted in Figure 2.11.

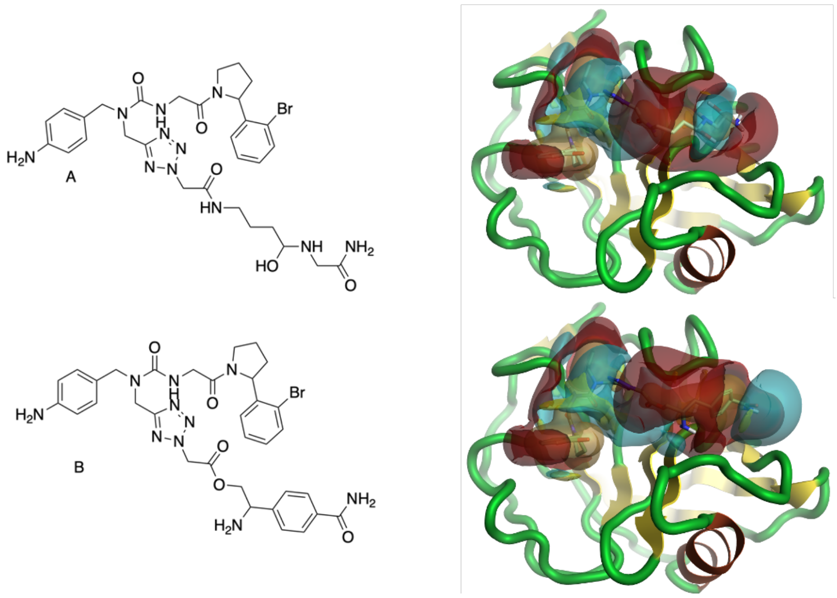


FIGURE 2.11: Chemical structures of the two *Reference* molecules, **A)** with linear substituents and **B)** with the 6-membered ring. In addition, the 3D representations of the toy molecules inside CypD are illustrated with their molecular fields. Red, blue, yellow and orange colours are used to show the negative, positive, vdW and hydrophobic surfaces respectively.

2.2.2 Selection of Compounds Based on Different Filters

For the selection of the desirable compounds from the virtual libraries to implement MD simulations, three filters were used: docking scores, synthetic feasibility and structural diversity. For the docking score calculations, all the compounds were docked to a CypD X-ray structure in complex with compound **15** reported by De Simone *et al*[156] using Cresset's molecular modelling package Flare[162]. The docking grid box was defined from compound **15** bound to CypD and the whole protein was used as a receptor. Lead Finder's algorithm was used as a scoring function.[165] It combines the classical genetic algorithm with multilevel optimisation procedures. Finally, only the lowest energetic pose from the type-II binding mode of each molecular structure has been chosen.

Addition filters were applied using two RDKit-based scripts.[166] RDKit is an open source toolkit that contains collection of machine learning and cheminformatics software written in C++ and Python and is used to develop custom applications for computer aided drug design. In our case, it provided the essential Python libraries (rdikit.Chem) for the molecular fingerprints needed for the application of the two filters. For the synthetic feasibility script, the SMARTS pattern was used to separate the solutions with N-C bond between the tetrazole and the fragment and the ones with N-N bond. We wanted to keep the former ones for the rest of the process.

The second RDKit-based script performed a clustering algorithm to identify structural diverse compounds. This script was applied to decrease the number of molecules that proceed to the next stages of the computational workflow that include computationally expensive MD simulations. The clustering was performed using the following procedure: the algorithm takes one compound and iteratively checks for similar compounds based on Tanimoto similarity and MACCS fingerprints. Once it finishes the search for the first compound and clusters the results together, it proceeds with the second molecule, etc. When this procedure finishes for every molecule of the dataset, the results are divided into two categories: *singleton* and *clustered*. *Singleton* contains all the compounds that were not structural similar with the rest of the dataset based on the applied Tanimoto similarity coefficient/value. *Clustered* consists off different subcategories of compounds with structural similarity based on the Tanimoto similarity value and MACCS fingerprints.

It should be noted that the choice of fingerprints has a major impact on the Tanimoto similarity value and that is the reason why MACCS fingerprints were used. MACCS keys consist of 116 bits recording 166 structural fragments. Each bit is associated with a SMARTS pattern. In addition, the Tanimoto similarity value was assessed with different tests using different cutoffs ranging from 0.75 to 0.98 as depicted in Figure 2.12. It was shown that Tanimoto similarity coefficient of 0.95 was optimal. Cutoffs below that threshold led to too few compounds retained, and above that threshold did not decrease significantly the size of the initial library.

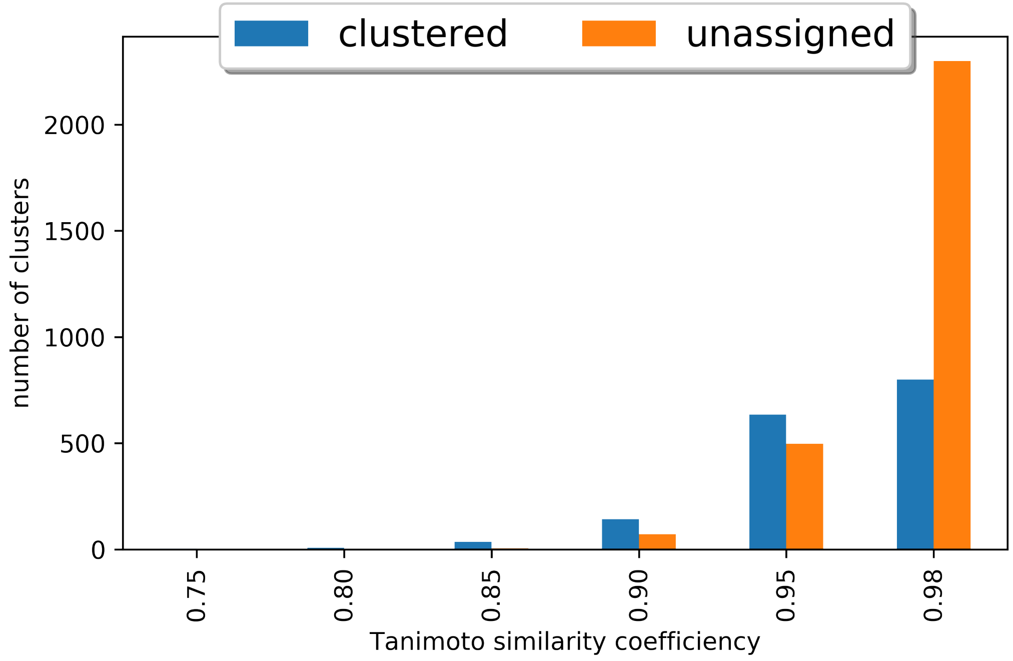


FIGURE 2.12: Assesment of the Tanimoto similarity value using different cutoffs. The Tanimoto similarity coefficiency is depicted in the x axis, while the number of clusters and the number of the unassigned compounds for each cutoff are shown in the y axis.

2.2.3 Molecular Dynamics Simulations of CypD-ligand complexes

For the Molecular Dynamics simulations, the binding pose from the previously performed docking calculations was chosen as a starting point for each structure. All the input files used for these simulations were created using FESetup1.2.1 software,[167] which is a python software for automated setup that uses AmberTools[168] for the parameterisation of protein-ligand complexes. Proteins were parameterised using ff14SB Amber force-field[169], while GAFF2 parameters[170],[171] that use AM1-BCC charges[172] were assigned to the ligands. All the protein-ligand complexes were solubilised in a rectangular box with TIP3P waters[173] with a box length whose edges extended 12 Å away from the edge of the solute. In addition, counter ions were added to neutralise the total net charge.

Next an equilibration protocol was performed, which included an energy minimisation of the system using 300 steepest descent steps followed by 700 steps of conjugate gradient steps in order to remove any possible artifacts from our system. Then, atoms in the solute molecules were position-restrained with a force constant of 10 kcal mol^{-1} Å^2 while a heating step to 300 K was performed for 200ps using an Andersen thermostat with a coupling constant of 10 ps^{-1}. Systems were then equilibrated for 1000 steps (2 fs timestep) using an NVT ensemble and the same restraints as in the previous step were used. Finally 5000 steps (2 fs timestep) of NPT ensemble at 1 atm (pressure control was maintained using a Monte Carlo barostat) were performed to reach a final density of about 1 g cm^{-3}. The final coordinate files were retrieved using the cpptraj module provided from AmberTools.

50 ns long MD simulations for every protein-ligand complex were run using the SOMD software (revision 2019.1.0) in the NPT ensemble at 300K and 1 atm. A 2 fs timestep was used and all the bonds involving hydrogens were constrained. Temperature control was maintained by an Andersen thermostat with a coupling constant of 10 ps^{-1}. Pressure control was achieved using a Monte Carlo barostat. Periodic boundary conditions were used with a 10 Å atom-based cutoff distance for the non-bonded interactions together with a Barker Watts reaction field with dielectric constant of 78.3 for the electrostatic interactions.

2.2.4 Relative Free Energy Calculations in CypA and CypD

Relative free energies of binding for all the suggested compounds were determined by alchemical free energy calculations.[23] For this purpose, a ligand is mutated into another in a water box, and in complex with CypA and CypD. Perturbation maps for compound series were generated by manual connection of the ligands via multiple transformations. Examples of perturbation maps are provided in the Appendix. The reference compound for all the perturbation maps was our lead compound 15 in complex with CypA and CypD. It should be noted that there is an offset of ca. 1 kcal mol^{-1} in absolute binding energies to CypD and CypA since compound **15** binds ca. 10-fold better to the first isoform. All the binding poses for CypD complexes were selected, parameterised and equilibrated as described in the Molecular Dynamics section 2.2.3. All the binding poses for CypA complexes were retrieved using docking calculations with Cresset's molecular modelling package Flare. The selected ligands from the MD simulations were docked to a CypA X-ray structure in complex with compound **15** reported by De Simone *et al*.[156] The aforementioned docking, preparation and equilibration protocols were used for these calculations.

Unless otherwise mentioned, all the simulations were run for 2 ns with SOMD in an NPT ensemble and the perturbed energies were saved every 250 fs. The number of equidistant λ windows employed for each perturbation was varied between 9, 17 or 26 (values for each window between 0.00 -1.00), based on the chemical similarity of the starting and the final compound. Before the production run, all the complexes were energy minimised for 1000 steps. A 2 fs timestep was used and a softcore potential was applied to keep pairwise interaction energies finite for all configurations and provide smooth free energy curves for all the simulations.[19] An Andersen thermostat and a Monte Carlo barostat were applied for the control of temperature and pressure respectively. Finally a 10 Å atom-based cutoff distance for the non-bonded interactions was used and Coulombic interactions were handled with a Barker Watts reaction field.

Free energy changes were estimated with the multistate Bennet acceptance ratio[50] as implemented in the Sire app *analysefreenrg* provided by SOMD. Convergence was assessed by checking the cycle closures in the perturbation maps (should be approx. 0) and the consistency between the free energies of binding from forward and backward simulations (the free energy difference for each forward and backward simulation should be within 1 kcal/mol). Simulations with poor convergence were repeated either with more lambda windows, if the perturbations had less than 26 lambda windows, or with more sampling time. The free energy analysis of the binding free energies of all the compounds relative to compound **15** was implemented through the *freenrgworkflow* python module.[174] Briefly, the free energies of binding from all the compounds were averaged for the forward and backward perturbations and the differences in free energies were then read into a Networkx (v 1.11) digraph, which is a Python language package for exploration and analysis of networks and network algorithms.[175] The estimation of the relative binding free energies of a given ligand to the reference compound 15 was performed by calculating all the paths connecting these two ligands. Then, the relative free energy of binding and its error estimate along each path was retrieved by adding the free energies along each edge of the path and by propagating the corresponding errors. Thus, the relative binding free energy between the two ligands can be calculated from the weighted average of all the unique paths such that more precise paths have a greater statistical weight.

2.3 Results

2.3.1 Construction of the virtual libraries and filtering of the compounds

A computational workflow was implemented for the discovery of second generation CypA and CypD selective inhibitors from our lead compound **15**. An overview of this protocol is depicted in the Figure 2.13.

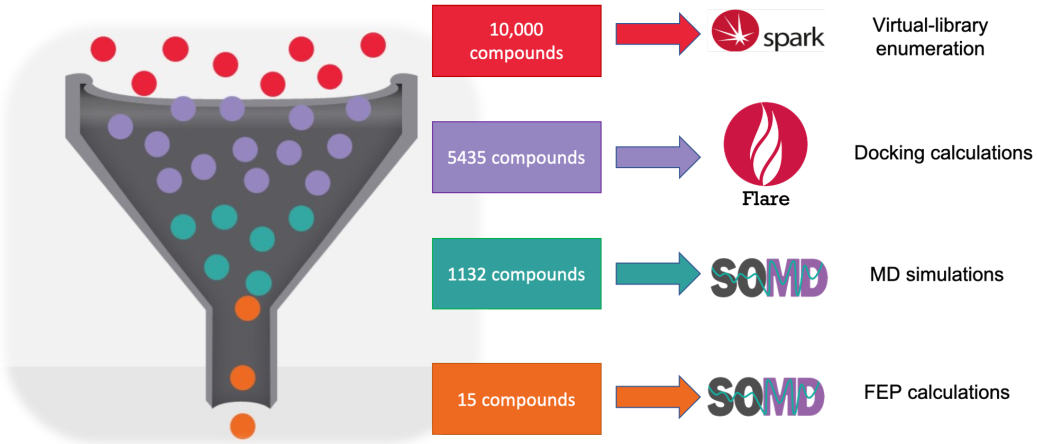


FIGURE 2.13: An overview of the computational workflow implemented for the discovery of selective CypA and CypD inhibitors. The procedure starts with a virtual-library enumeration of 10.000 compounds and continues with docking and similarity clustering of these compounds to produce a diverse subset of 1132 ligands in binding mode II. Then MD simulations are employed to retain only 15 stably bound analogues. Finally, FEP calculations are applied to select compounds with potency and selectivity improvements. The resulting compounds will be synthesised and characterised by biological assays.

For the first step of this protocol, two virtual libraries were constructed each with 5.000 different analogues of compound 15, in order to explore the chemical space for binding to the *3 o'clock* pocket. All the compounds were docked using Flare in order to acquire the lowest energetic structure in the type-II binding mode for each structure. Docking was essential, as Spark is a ligand-based method for the creation of the virtual libraries. Therefore, it did not provide evidence about the energetic stability of the conformations of the product molecules in type-II binding mode. The desired binding mode for each compound is illustrated in Figure 2.14:

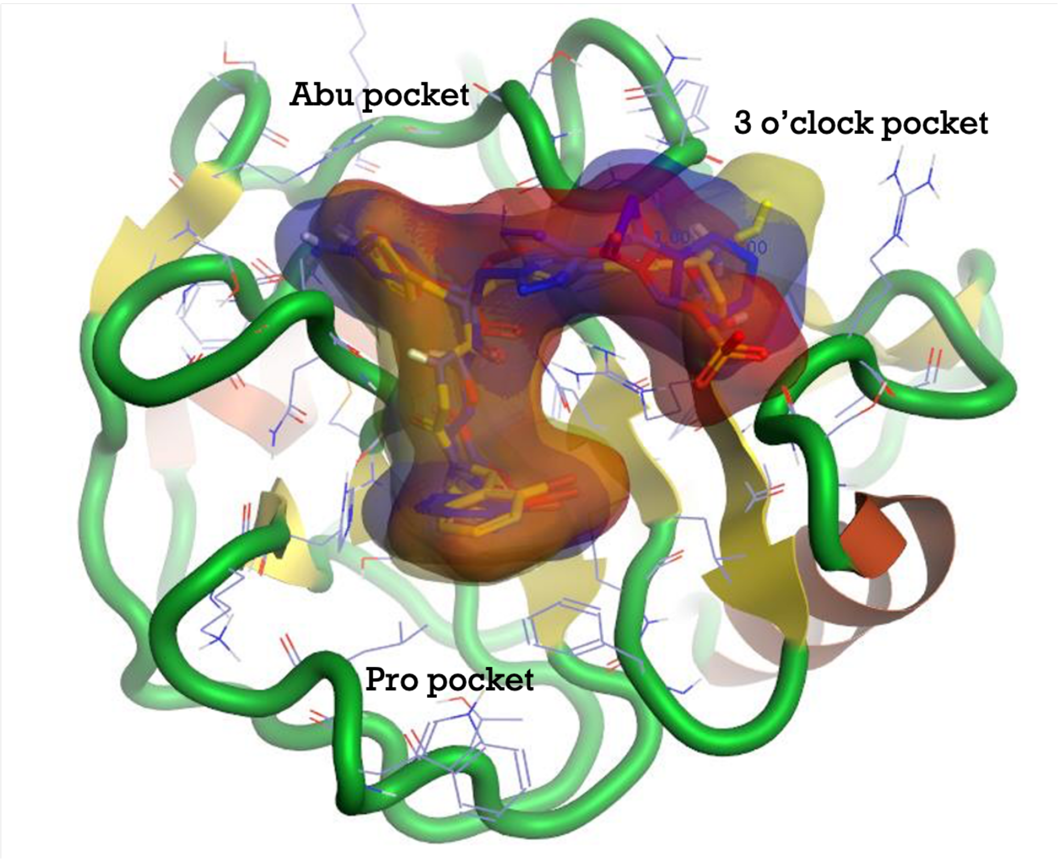


FIGURE 2.14: Depiction of the desired binding mode in CypD. Three docked ligands are shown with different colors, red, blue and yellow in type-II binding mode.

The compounds in binding mode II were then filtered according to their synthetic feasibility. Solutions that featured a N-N bond between the tetrazole and the substituent were excluded. Such bond could be labile as the bond dissociation energy of two single bonded heteroatoms is typically lower than for a single bond between one carbon atom and one heteroatom. This decreased the virtual library size down to 5435 analogues. The distribution of the docking scores of these molecules is depicted in Figure 2.15.

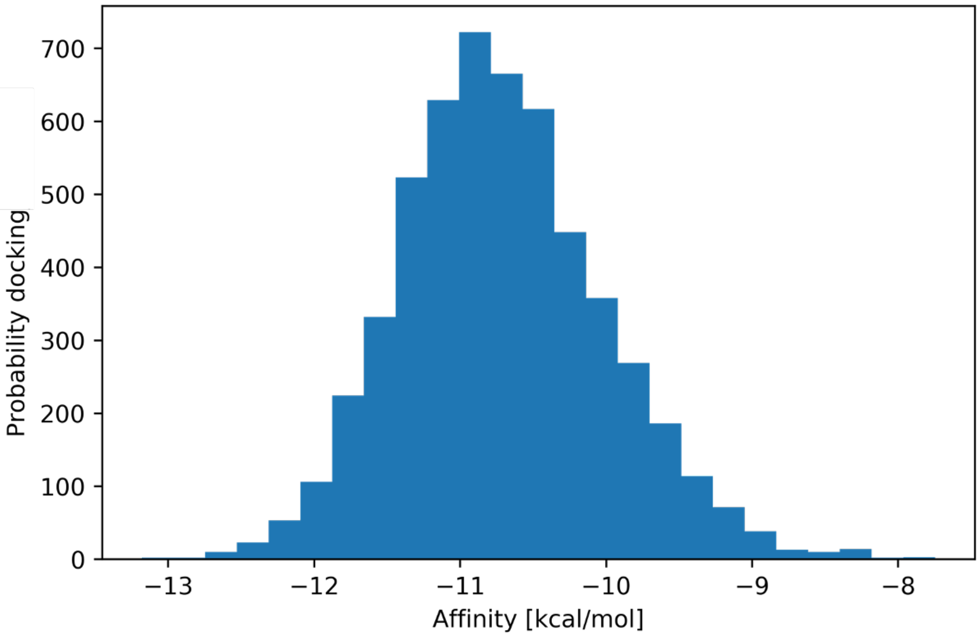


FIGURE 2.15: Distribution of docking scores for 5435 compounds with a carbon atom bonded to the tetrazole ring. The mean value of the docking scores is around -11.00 kcal/mol. The docking score of compound **15** is equal to -10.30 kcal/-mol.

The remaining compounds were further refined based on their structural similarity. The goal was to limit the number of compounds that proceed to the next step of the protocol that features expensive MD simulations. For this purpose, the molecules were divided into two categories, *singleton* and *clustered* based on their MACCS fingerprints and a Tanimoto similarity coefficiency of 0.95. The most representative structure of each cluster together with the *singleton* compounds were chosen for the next stage of our protocol. This filter limited the number of promising compounds to 1132.

2.3.2 Molecular Dynamics results

Further refinement of compound prioritisation was accomplished by means of MD simulations. The goal was to retain only compounds that maintain stable interactions within the *3 o'clock* pocket. For this reason, 50 ns long MD simulations were performed for each of the remaining 1132 compounds in complex with CypD. The first criterion to define a stable interaction during the course of the MD simulation was the calculation of the RMSD of the -R groups connected to tetrazole inside the *3 o'clock* pocket. The overall results from the RMSD filter are illustrated in Figure 2.16.

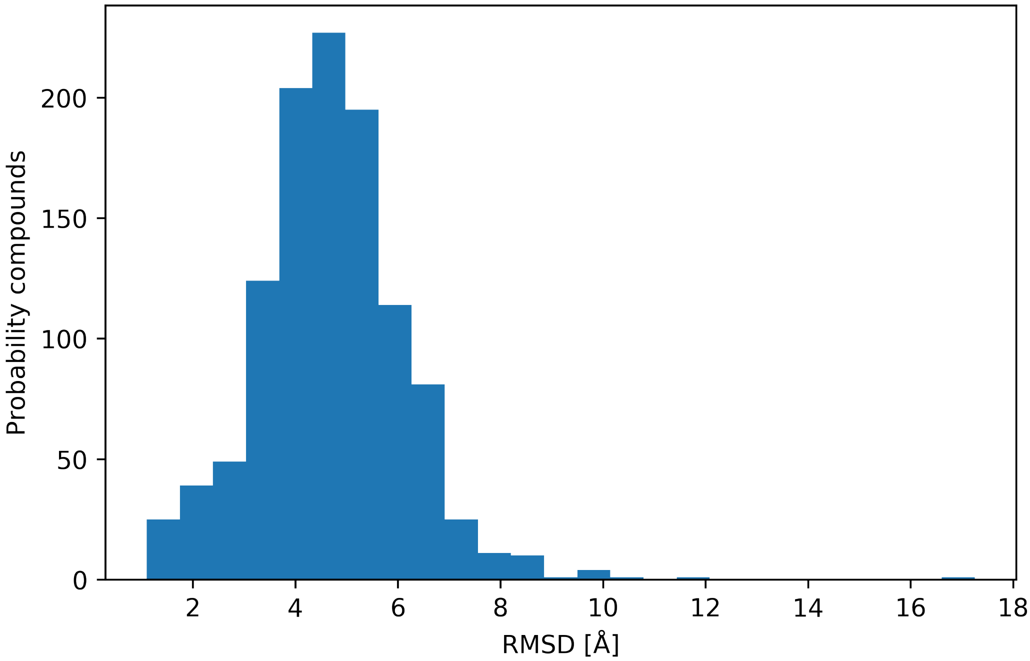


FIGURE 2.16: Depiction of the histogram of RMSD values in the 1132 analogues of compound **15**.

It was observed that only 72 compounds exhibit RMSD values below 2 Å. This threshold was chosen as compounds with higher RMSD values tended to show only weak or transient interactions with residues in the *3 o'clock* pocket.

Two additional filters were applied to identify compounds predicted to bind more favourably to CypD. The preservation of the hydrogen bonds that are essential for a type-II binding to CypD was the first filter. For this purpose, a 50-ns MD simulation was performed for the aryl pyrrolidine compound with Hydrogen instead of tetrazole in a type II binding mode. The percentage of time hydrogen bonds are formed between this compound and key residues in *Abu* and *Pro* pocket during the simulation (Thr107, Asn102, Gln63 and Arg55) were computed using cpptraj. The same analysis was performed for the 72 CypD-ligand complexes and these numbers were subtracted from the percentages of the reference compound. These differences were summed up to provide a final number. In this study, this number is called *final hydrogen bond value*: A positive number is associated with compounds that form longer-lived hydrogen bonds than the reference compound. The results for the 72 compounds are summarised in Figure 2.17

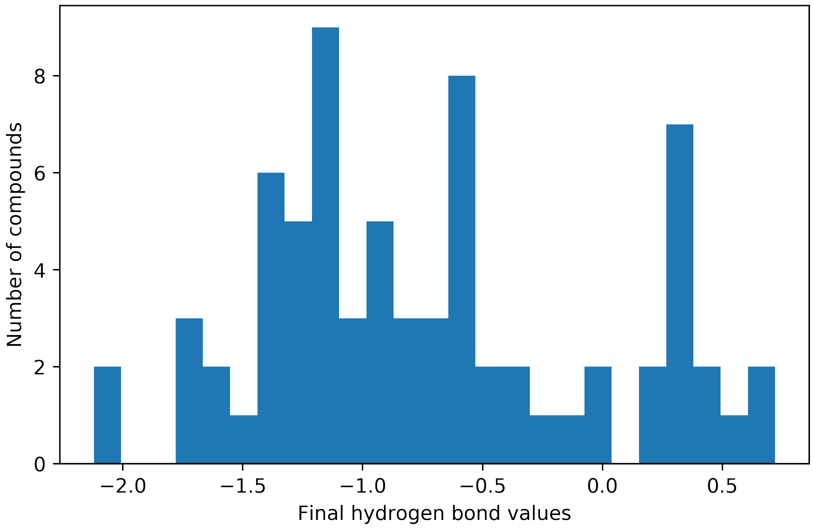


FIGURE 2.17: Depiction of the histogram of the final hydrogen bond values in the 72 analogues of compound **15**.

The hydrogen bond formation between the R-groups of the 72 analogues of compound 15 and *3 o'clock* pocket residues was also examined with the same procedure used for *Abu* and *Pro* pocket residues. The aim of this filter was to identify R groups that prefer to stay inside the *3 o'clock* pocket and it resulted to 23 compounds that fulfilled this criterion. The combination of this filter together with the essential hydrogen bonds for a type-II binding mode, resulted in 26 compounds that were identified as the most promising solutions for CypD selectivity. The chemical structure of the 26 compounds that fulfilled one or both of the aforementioned criteria are depicted in Figure 2.18.

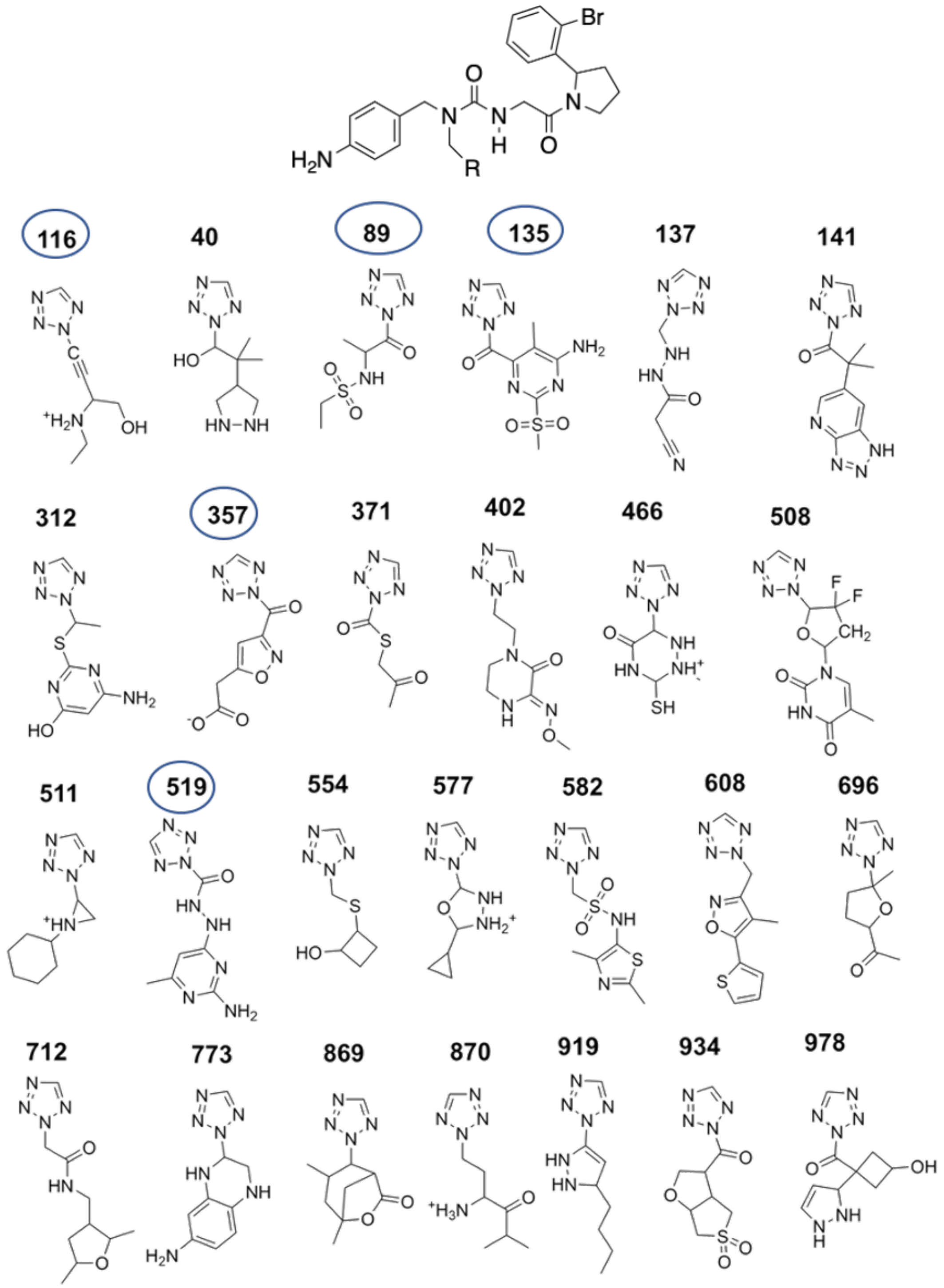


FIGURE 2.18: Structures of the 26 analogues of compound **15** that meet one or two of the two following criteria: stable type II binding mode, stable interactions in the *3 o'clock* pocket. The compounds chosen for FEP calculations after visual inspection of the MD simulations are shown in blue circles.

Visual inspection of the trajectories of those compounds using the software VMD was performed to examine the behavior of the R groups. The purpose of this test was to ensure that the R groups were staying inside the *3 o′clock* pocket and they were not solvent exposed. This led to selection of five analogues for further assessment using FEP calculations.

Compound **116** was the best molecule amongst the five most promising designs, as it maintained the interactions within the *Abu* and *Pro* pocket, formed strong hydrogen bonds inside the *3 o′clock* pocket and visually demonstrated an appealing conformation inside the accessory pocket via its tetrazole-alkyne motif. The other four compounds, **89**, **135**, **357** and **519** had all in common the tetrazole-ketone motif that maintaining stable interactions with *3 o′clock* pocket residues, with the possible exception of compound **89**. Compound **89** was chosen for its strong interactions with key Abu/Pro residues, while compounds **357** and **519** did not form hydrogen bonds with *Abu* and *Pro* pocket residues for a long period of time. Compound **135** was the only ligand from these series that formed strong hydrogen bonds inside all three key pockets.

Since a tetrazole-ketone motif emerged as an interesting design choice, the conformational preferences of ligands featuring this group were examined with a torsional scan (24 scanning windows). The results depicted in Figure 2.19 were obtained via Density-functional theory (DFT) calculations using the software Gaussian.[**g16**] The examined compound was simulated in gas state using the B3LYP/6-31G basis set and a 10^{-6} Hartree convergence criteria. The results provided further evidence that the conformation observed during the MD simulations is energetically low, and thus favorable for binding.

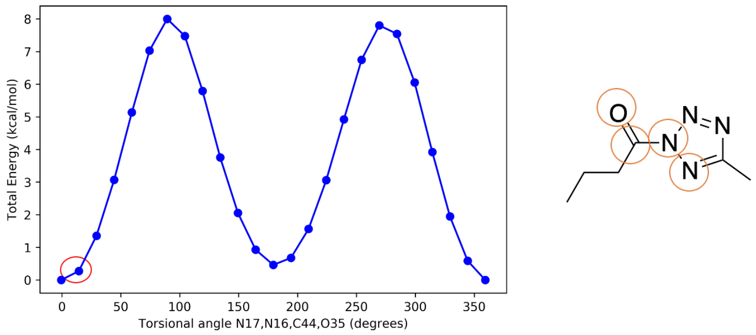


FIGURE 2.19: Torsional scan of the tetrazole-ketone dihedral angle that includes the 4 atoms as they are depicted in the 2D representation of the toy molecule on the right picture. The degrees of the dihedral angle during the torsional scan are depicted on the x-axis and the energy in kcal/mol on the y axis. The average value of the dihedral angle in the tetrazole-ketone motif observed during MD simulations of 89 is highlighted with a red circle.

Finally, analogues of the five chosen compounds were also explored to identify molecules more likely to bind selectively to CypD over CypA. They were based on the tetrazole-alkyne motif and the tetrazole-ketone motif described above. For the former motif, three analogues of compound **116** were chosen to replace the second chiral center introduced from the flexible chain after the alkyne to avoid introduction of another chiral center. The amine group of the chain was removed from the first molecule, while the more stable morpholine and piperidine rings where introduced. In addition, bioisosteric replacements of the alkyne-flexible chain motif were also suggested by using the corresponding feature in the Spark software. One compound suggested for FEP calculations, that contained a disulfide bond, showed in the MD simulations stable interactions with the *Abu*, *Pro* and *3 o'clock* pockets. Ultimately, the replacement of the alkyne motif with oxazole provided a ligand with very stable interactions in the *Abu* and *Pro* pockets.

For the tetrazole-ketone motif, different analogues were designed to explore different 6-membered ring substitution patterns, as well as linear and 5-membered ring alternatives. All the compounds chosen from the MD simulations for the relative FEP calculations are illustrated in Figure 2.20.

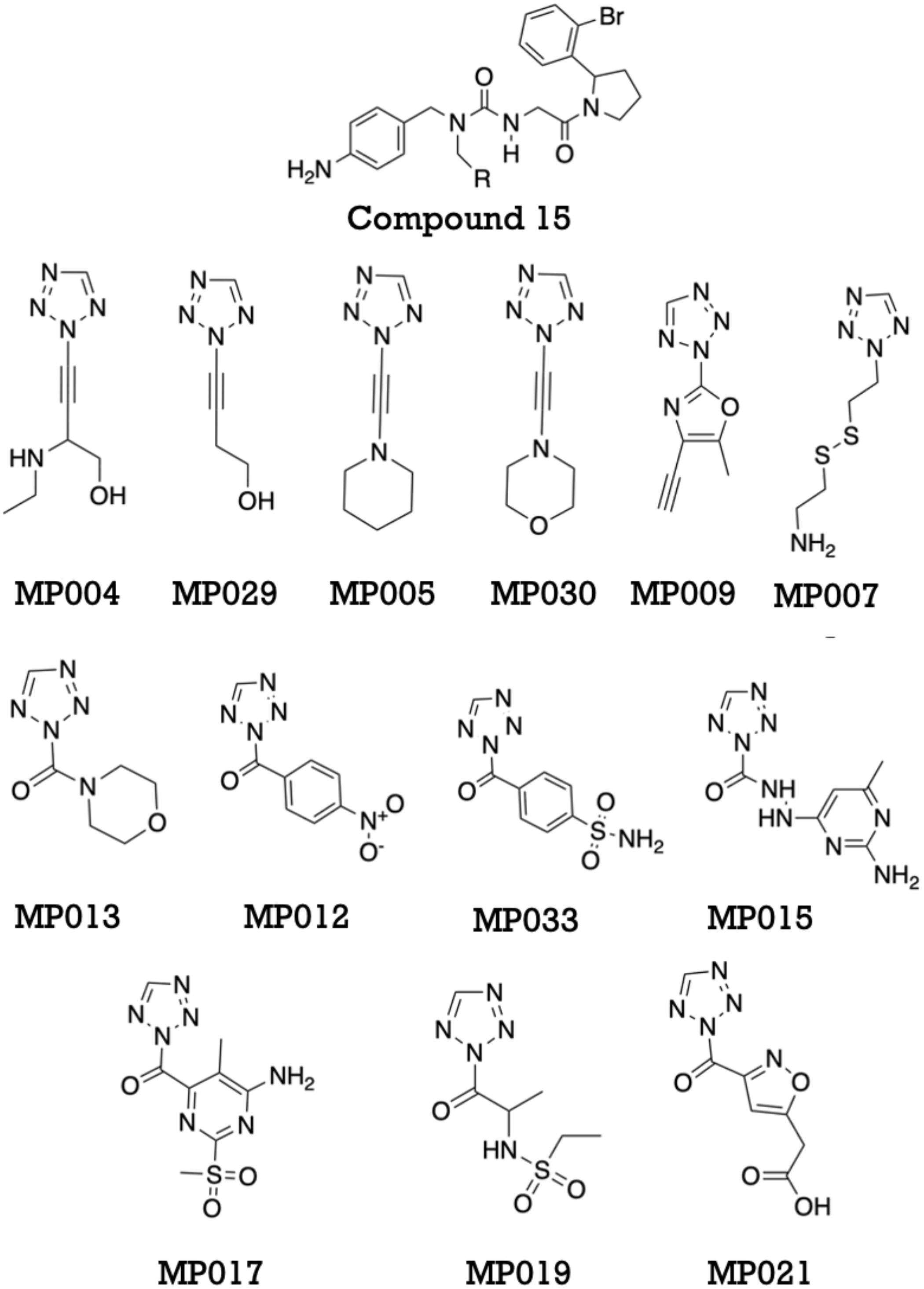


FIGURE 2.20: The most promising solutions for improved potency and selectivity compared to compound **15** as chosen from the MD simulations.

2.3.3 Free Energy Perturbations Results

The most promising solutions were further assessed using Free Energy Perturbation to identify designs more likely to bind selectively to CypD over CypA. For the tetrazole-alkyne motif, all the analogues showed improvement in potency compared to compound **15**. In addition, the morpholine group showed also better selectivity (ca. 1 kcal/mol for CypD over CypA) and this compound was deemed the most promising in this series of compounds. Moreover, the perturbation network of the disulfide derivatives was further examined to evaluate the role of -NH2 group in binding potency and selectivity, as these analogues showed preference for CypA over CypD. For this purpose, a more extended perturbation network was designed with ligands that did not contain the tested amine group.

The results from the relative binding free energies did not show energetic preference for the -NH2 group over the heptane analogues. To further prove this point, the hydrogen bonds formed by MP006 and MP007 at the end state of the bound vanish step were analysed through cpptraj and provided also evidence that NH2 was not the key factor for the strong binding to Cyps. The overall results from the tetrazole-alkyne perturbation networks are illustrated in Figure 2.21

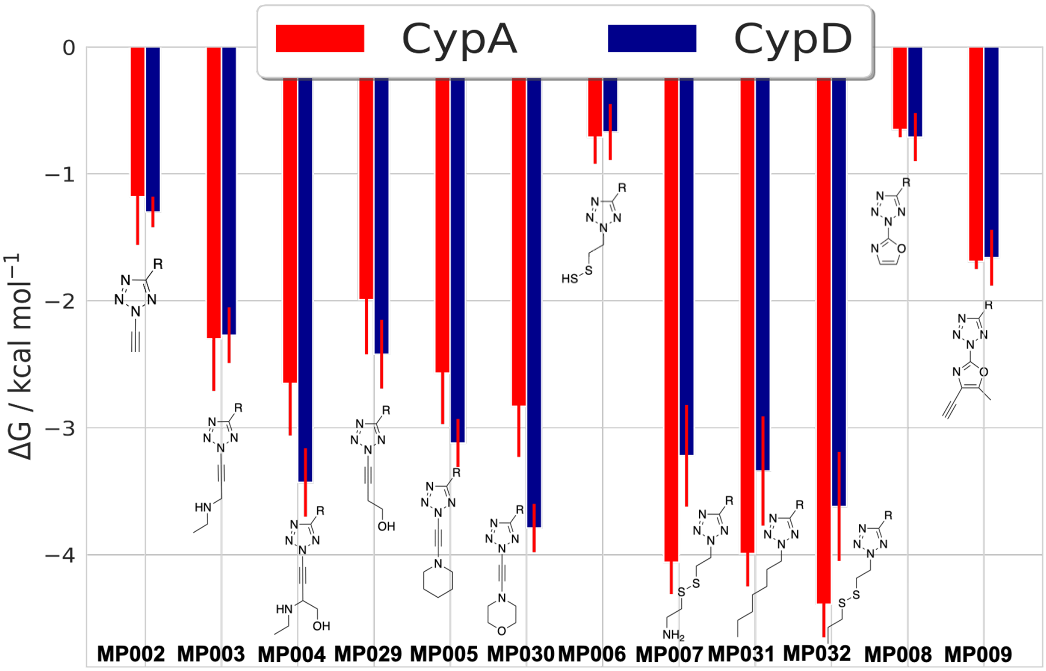


FIGURE 2.21: Computed relative binding energies for tetrazole-alkyne analogues.

For the tetrazole-ketone motif, all the analogues of compound **15** predicted to bind more strongly to CypD and CypA compared to the parent compound, with the possible exception of MP015. The most auspicious ligands from these series were MP012 and MP033 that had in common the para-benzene group. Both compounds were exhibited a 2 kcal/mol energetic preference for CypA over CypD. Based on these encouraging results, the tetrazole ring of those two compounds was replaced with an oxadiazole ring as such analogues were deemed potentially more synthetically feasible. The new compounds also retained the potency and selectivity of the tetrazole derivatives and were identified as the most promising designs of this series. The overall results from the tetrazole-alkyne perturbation networks are illustrated in Figure 2.22

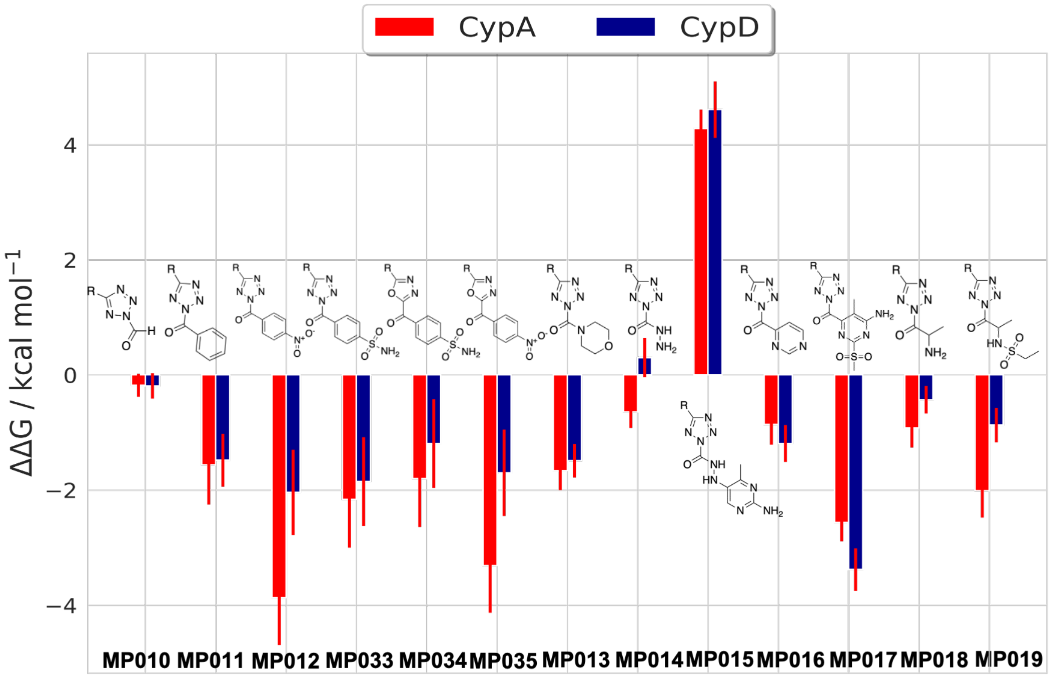


FIGURE 2.22: Computed binding energies for analogues featuring a tetrazole-ketone motif.

Increased emphasis was given to the MP020-MP021 perturbation network due to the results of the MD simulations, which showed that the acid was contributing to the stability of the fragment inside the *3 o'clock* pocket. FEP calculations suggested that MP021 improves the potency in Cyps by 2 kcal/mol compared to compound **15**. Different substituents on the isoxazole ring were used to examine the importance of the acid spanning from a linear propane to an amide. All of the analogues showed the same energetic preference for the Cyps as with MP021, suggesting that the isoxazole scaffold is the most important feature of these series of compounds. Ultimately, the replacement of the tetrazole ring with the oxadiazole ring was also examined. The relative binding free energies of the oxadiazole ligands maintained the enhancement in binding that is observed for this motif. Thus, MP021 and MP041 are the most promising solutions from this batch of promising compounds. The overall results from the MP020-MP021 perturbation networks are illustrated in Figure 2.23

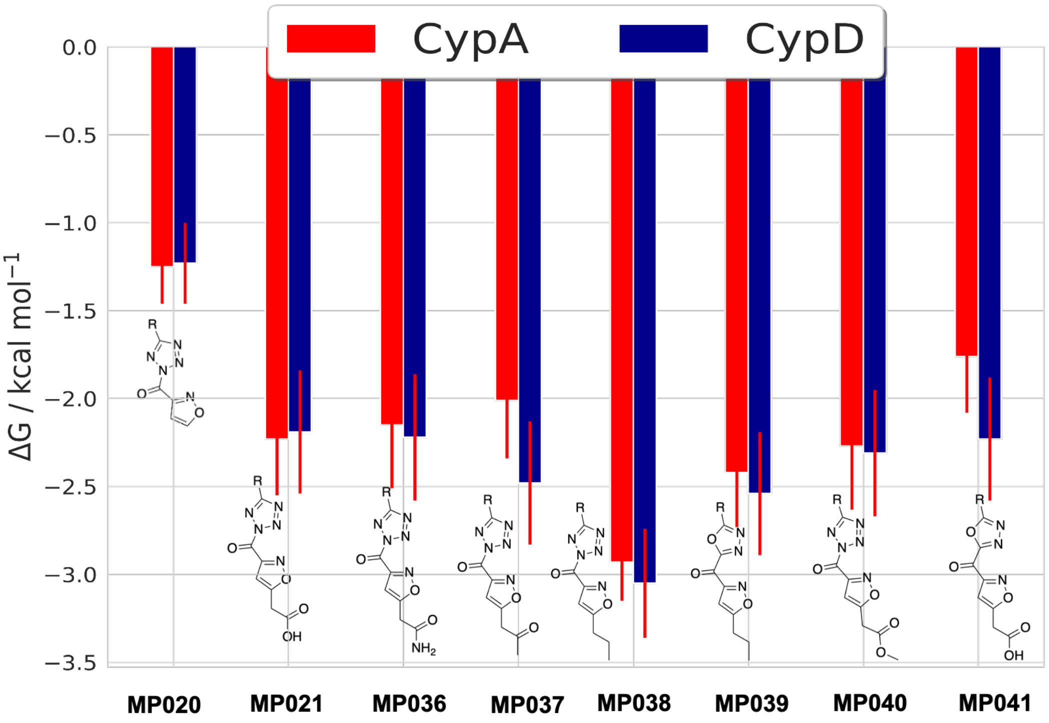


FIGURE 2.23: Computed binding energies for ketone-oxazole series.

Based on the first promising solutions, a second round of relative FEP calculations was undertaken to explore new scaffolds for CypD or CypA selectivity. The replacement of the morpholine analogue with a tetrahydropyran ring together with the addition of the cyclopropyl group in the current motifs were the main targets in this batch of simulations. The latter adjustment did not favor binding in CypA and CypD compared to compound **15**. However, the tetrahydropyran ligand offered a ca. 2 kcal/mol preference in binding for CypD over CypA, together with a 4 kcal/mol improvement in binding energy over the parent compound. Taking this into consideration, a visual inspection at the end state of the bound discharge step of the tetrahydropyran perturbation was performed to provide evidence for this selectivity. It was observed that in CypA the tetrahydropyran analogue was solvent exposed after 40 % of the simulation agreeing with the difference in binding free energy between the 2 proteins. In addition, we expanded the oxadiazole ring instead of the tetrazole in more compounds, including compound

15, to have a more thorough investigation for the preserving of binding to Cyps. This point was proved also from the second batch of simulations making these compounds perfect candidates for synthesis and biophysical characterisation. The overall results from the second batch of simulations are depicted in Figure 2.24

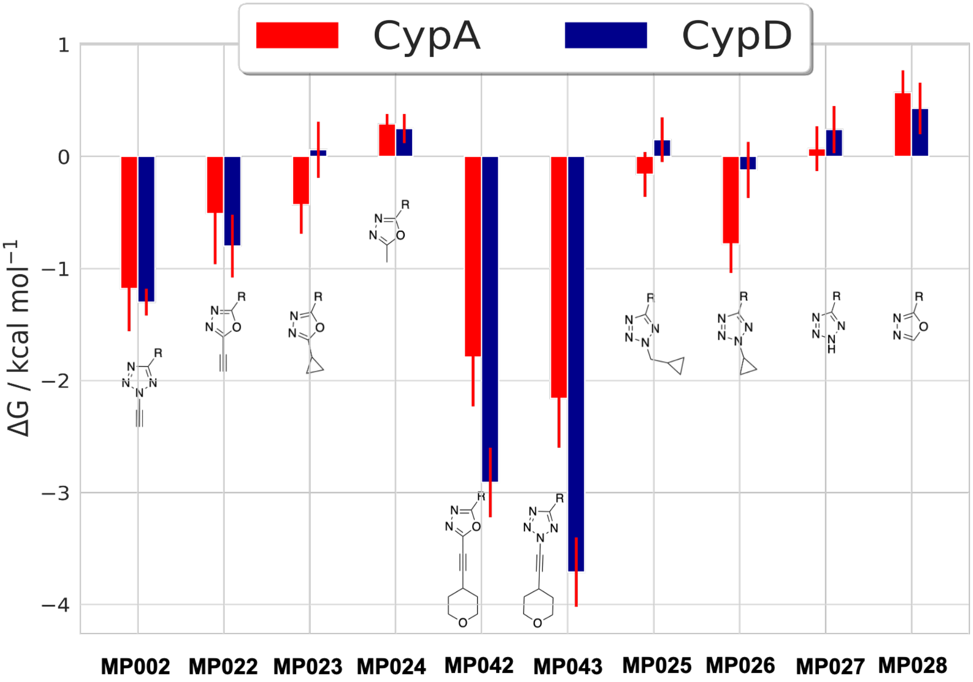


FIGURE 2.24: Computed binding energies for the second batch of simulations in CypA and CypD.

Next, given the feasibility of replacing the tetrazole by an oxadiazole ring, a range of other rings were examined through relative FEP calculations. Alternatives to a tetrazole ring that reduce the number of hydrogen bond acceptors could improve Blood Brain Barrier penetration of tri-vector cyclophilin compounds. For this purpose, a perturbation network was devised between compound **15** and other ring replacements such as furane and thiophene. The results shown in figure 2.25 provided evidence that such replacements would generally weaken binding with respect to compound 15, albeit not excessively.

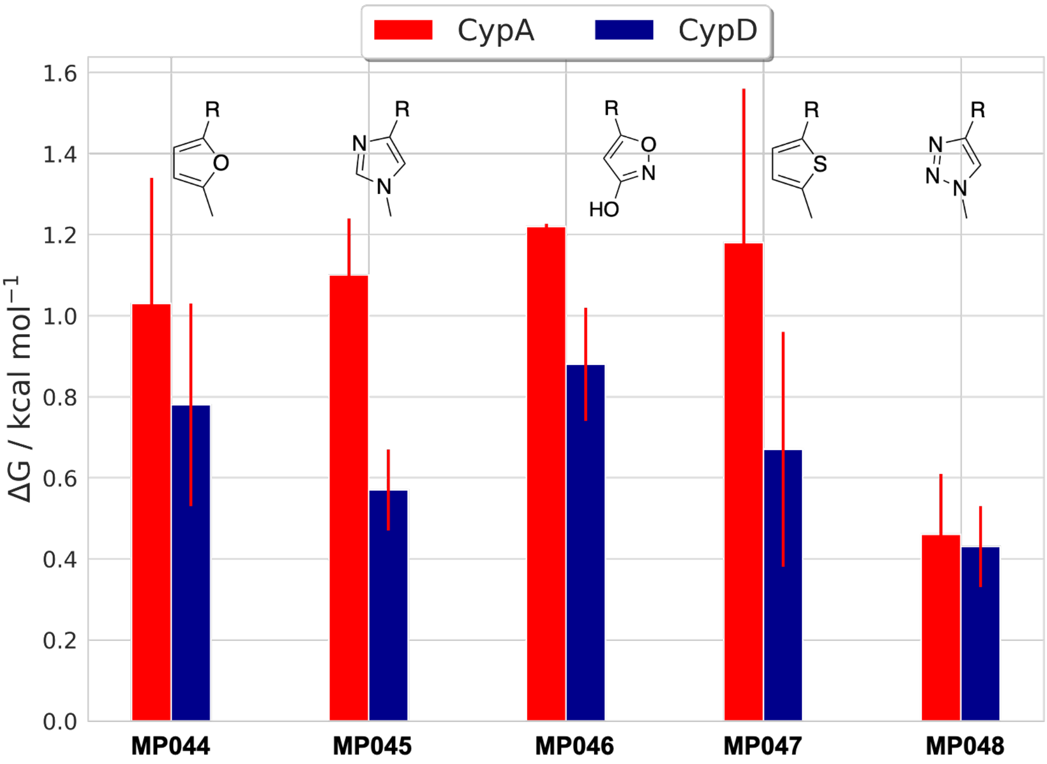
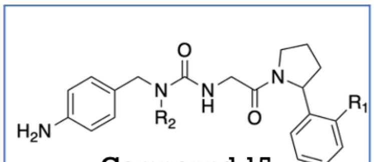


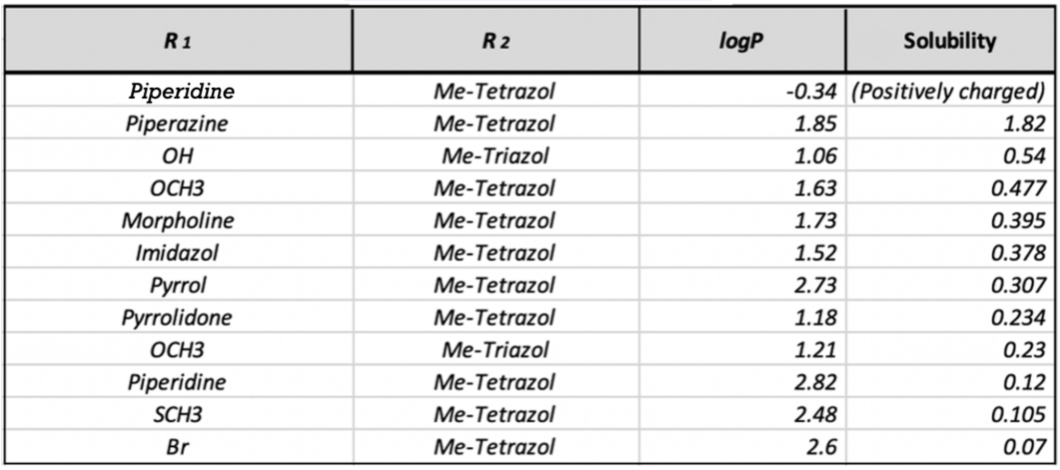
FIGURE 2.25: Illustration of the free energies of binding of different ring replacements compared to compound 15 in CypD.

2.3.4 Pro pocket results

The main goal of the Pro pocket optimisation was to replace the bromine atom in the arylpyrrolidine motif with a substituent to improve solubility. For this purpose, additional FEP calculations were carried out on a set of compounds with predicted improved solubility based on calculations carried out by Dr. Jordi Juárez Jiménez using the software ChemAxon.[176]



Compound 15



R 1	R 2	logP	Solubility
Piperidine	Me-Tetrazol	-0.34	(Positively charged)
Piperazine	Me-Tetrazol	1.85	1.82
OH	Me-Triazol	1.06	0.54
OCH3	Me-Tetrazol	1.63	0.477
Morpholine	Me-Tetrazol	1.73	0.395
Imidazol	Me-Tetrazol	1.52	0.378
Pyrrol	Me-Tetrazol	2.73	0.307
Pyrrolidone	Me-Tetrazol	1.18	0.234
OCH3	Me-Triazol	1.21	0.23
Piperidine	Me-Tetrazol	2.82	0.12
SCH3	Me-Tetrazol	2.48	0.105
Br	Me-Tetrazol	2.6	0.07

FIGURE 2.26: Predicted LogP and solubility for analogues of compound **15**.

Based on these calculations, all the substituents were predicted to improve the solubility of the parent compound. However, most of the ligands of this series of compounds tend to slightly worsen the binding of our lead molecule. Only piperidine and morpholine derivatives tend to maintain the potency and are promising designs for targeting the *Pro* pocket. Curiously removal of ortho-substituents seems to offer 1 kcal/mol selectivity for CypA over CypD. The overall results of these calculations are summarised in Figure 2.27.

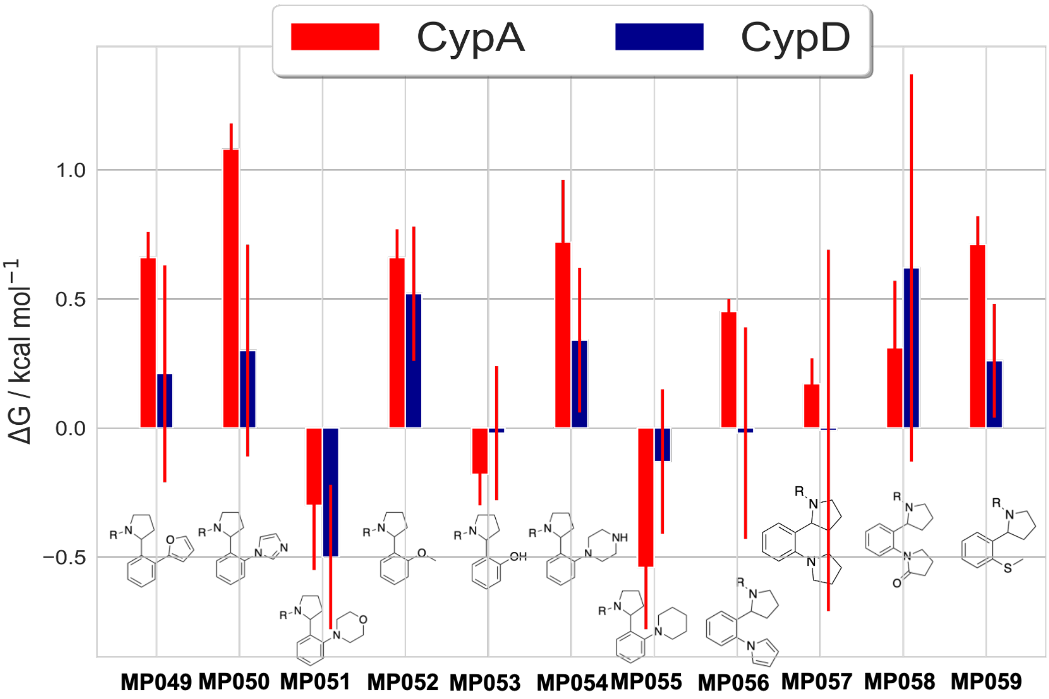


FIGURE 2.27: Depiction of the free energies of binding of the *Pro* pocket alternatives to compound 15 in CypA and CypD.

2.3.5 Abu pocket results

The last aim of this computational study was to reduce risks of genotoxicity in the lead molecule through modifications of the aniline ring that targets the *Abu* pocket. Modification at sites ortho to the amino group were sought to block potential oxidation sites that could lead to the formation of toxic metabolites. For this purpose, two potential candidates were suggested to retain the potency and reduce the toxicity of compound **15**. Relative FEP calculations were also performed for these molecules in CypA and CypD compared to the parent compound.

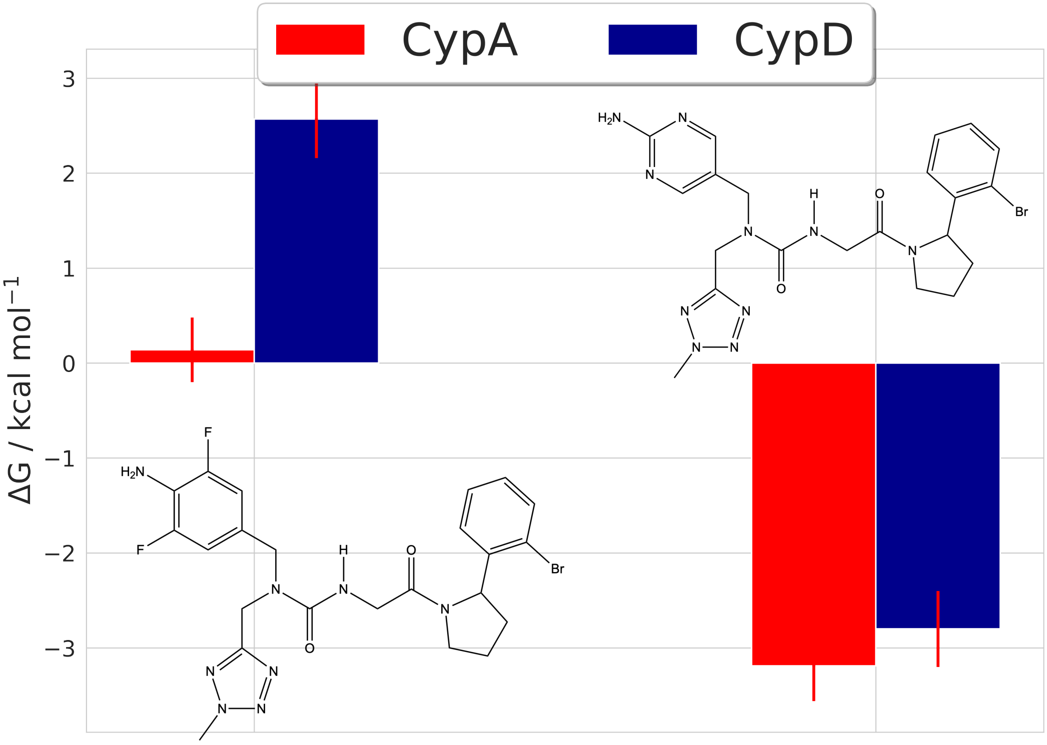


FIGURE 2.28: Illustration of the free energies of binding of the *Abu* pocket substituents compared to compound **15** in CypA and CypD.

The results from these calculations showed that the compound with the fluorines in ortho position did not retain binding in CypD but offered selectivity of 2 kcal/mol for CypA over CypD. However, visualisation of the average structure of this ligand in CypA and CypD adopted during the FEP simulations showed that the ligand of interest was solvent exposed in the *Abu* pocket of CypA. Therefore, this result was at odd with the calculated energetics. Thus, a follow up 100 ns MD simulations of the compound of interest was performed in CypA and in CypD. The visualisation of the trajectories in both Cyclophilins justified the energetic preference for this ligand in CypA over CypD as it was more stable in binding mode II. This conclusion was also confirmed after calculating the formation of the hydrogen bonds of the compound of interest with both Cyps. The binding of the 2,6-fluorine analogues to CypA and CypD is shown in Figure 2.29.

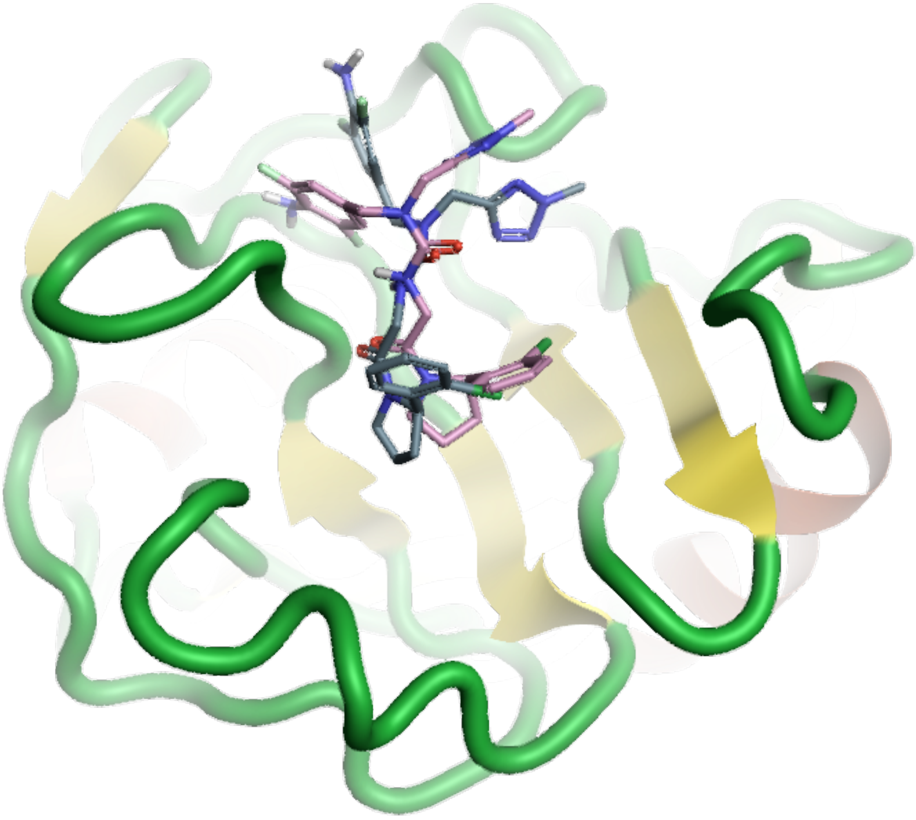


FIGURE 2.29: Depiction of 2,6-fluorine analogues bound to CypA (grey) and CypD (pink).

Two 100 ns MD simulation in CypA and CypD were also performed for the pyrimidine analogue, as it showed more favorable binding to the proteins of interest than the lead compound. This molecule had a longer hydrogen bond formation with the key residue for binding in *Abu* pocket, Thr107, as well as a more stable type-II binding mode compared to the 100 ns MD simulations of compound **15** in CypA and CypD. Therefore, it could be an interesting design for addressing the toxicity issue and also increase potency.

A third analogue of compound **15** was also examined based recent work by Grädler *et al* that synthesized a series of linear (type I binding mode) CypD inhibitors. In this paper the most potent derivative, compound **2** contains a bicyclic fragment instead of an aniline ring. Therefore, we evaluated whether this bicyclic fragment could replace the aniline ring in the context of a type II binding mode compound (Figure 2.30). For this purpose, 100 ns MD simulations in CypA and CypD were performed to assess the stability of the designs. The bicyclic derivative was very stable in both proteins making it a suitable candidate for the replacement of the aniline ring.[177]

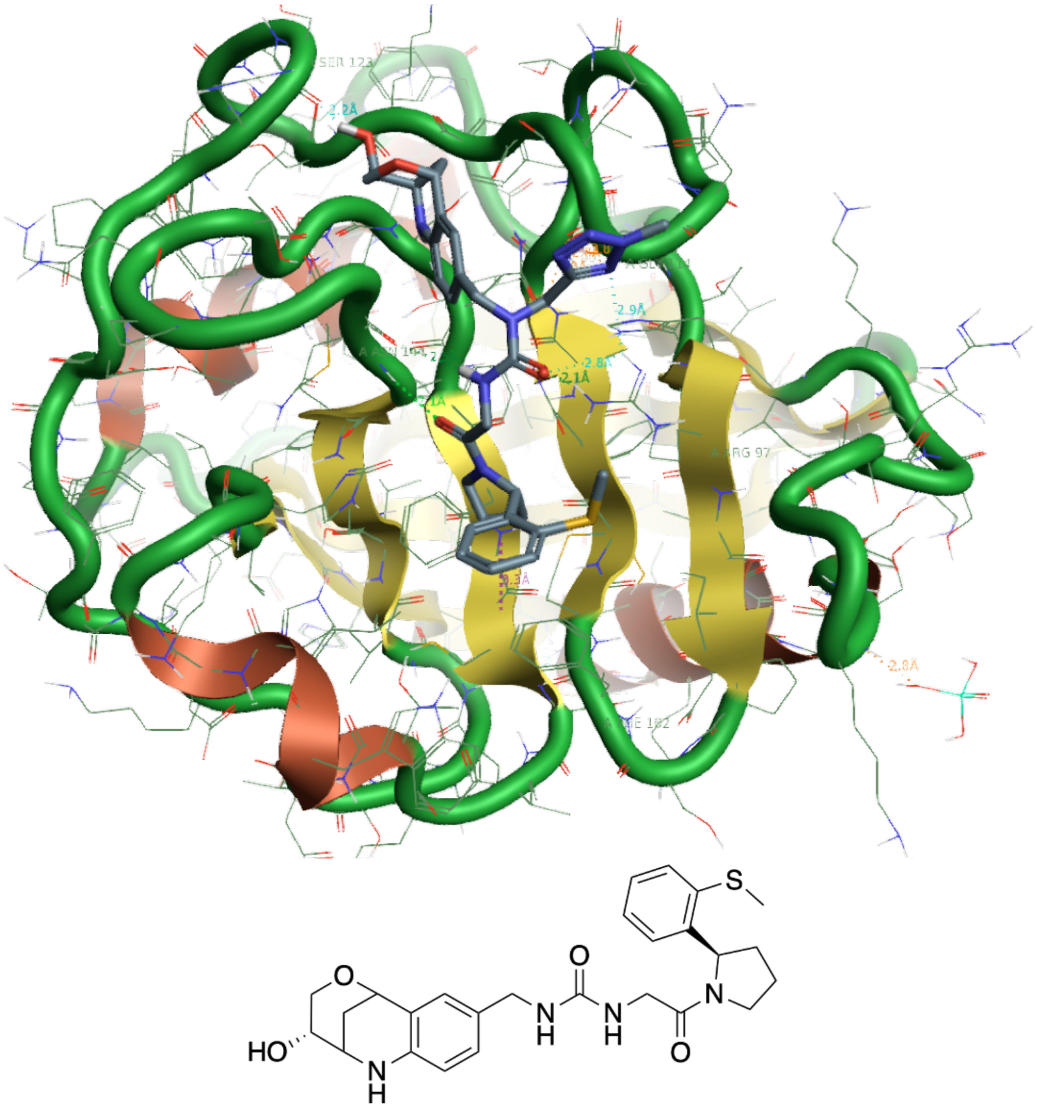


FIGURE 2.30: Illustration of the 3D representation of the bicyclic analogue of compound **15** in CypA using Flare. In addition, the chemical structure of the bicyclic analogue is also depicted.

2.4 Conclusions

A novel computational workflow was implemented for the discovery of selective CypA and CypD inhibitors by modifying compound 15 to protrude deeper into the *3 o'clock* pocket. To explore the chemical space for binding to this pocket, a library of ca. 10000 analogues of this molecule was constructed and docked using the Cresset molecular modelling package. Compounds were filtered according to their predicted synthetic feasibility, clustered into structurally diverse families and ranked using the docking scores. Further refinement of compound prioritisation was accomplished by means of MD simulations, aiming to retain only compounds that maintain stable interactions within the *3 o'clock* pocket. The most promising solutions were further assessed using Free Energy Perturbation calculations to identify designs more likely to bind selectively to one of the two isoforms.

Based on this workflow different scaffolds were discovered that could lead to promising solutions for improved potency and selectivity to the proteins of interest. Both the tetrazole-alkyne and tetrazole-ketone motifs enhanced the binding of our lead compound to CypA and CypD. The particular highlights of the tetrazole-alkyne series were MP030, that showed 1 kcal/mol energetic preference to CypD over CypA, the tetrahydropyran analogue that offers substantial binding and selectivity for CypD over CypA, and the disulfide derivatives that were predicted to bind more strongly to CypA over CypD. The most encouraging designs from the tetrazole-ketone motif were the para-benzene and the oxazole analogues that offered significant binding to cyclophilins compared to our lead compound ranging from 2-4 kcal/mol energetic preference. Different rings were examined to replace the tetrazole for the improvement of the physicochemical properties and the synthetic feasibility of the second generation inhibitors of Cyps. The most promising result was the oxadiazole derivative that could ease the synthesis of these designs while maintaining the potency and selectivity trends.

Moreover in this study, MD/FEP methods were able to predict scaffolds that could improve the physicochemical properties of our lead compound. The compounds that are promising candidates for biophysical characterisation are only the compounds where bromine was replaced by piperidine and morpholine. Ultimately, this study also shed light on compounds with reduced toxicity compared to our lead compound. A bicyclic derivative from Grädler *et al* was predicted to be very stable in both Cyps after 100 ns of MD simulations and a pyrimidine analogue showed a 3 kcal/mol energetic preference for the proteins of interest from relative FEP calculations.[177]

The results from this work have informed the synthesis of second generation tri-vector inhibitors. This will be followed by characterisation of the most promising compounds using biophysical assays. Therefore, in the near future, a complete study can be performed in the quest for a novel class of selective Cyclophilin inhibitors and may thus the whole strategy can be conducted across the board of structure-based drug design.

Chapter 3

Blinded Predictions of Standard Binding Free Energies: Lessons Learned from the SAMPL6 Challenge

3.1 Introduction

As mentioned in Chapter 1, the accurate prediction of protein-ligand binding free energies is a principal target of computer-aided drug design (CADD). The precise description of ligand-protein energetics, is nowadays increasingly sought via use of free energy calculations methods. Among many existing free energy calculation methodologies, alchemical free energy (AFE) calculations have attracted much interest in recent years,[174, 178, 179] due to their strong grounding in statistical physics. AFE calculations capture non-additivity of structure-activity relationship in congeneric series that are overlooked by empirical scoring methods[180], and have given useful potency predictions for a plethora of protein-ligand systems.[81, 181, 182]In addition, AFE methods may be used to predict physical properties, such as lipophilicity coefficients.[183–185] In spite of encouraging successes, there are still important technical hurdles to tackle. Usual concerns involve finite-sampling effects that introduce statistical errors,[186–189] while inaccuracies in potential energy functions contribute to systematic errors.[190] Additionally, decisions of the appropriate algorithms handling the long range electrostatic interactions and finite-size artefacts, affect simulation results in ways that are still poorly understood, with effects particularly apparent in the modelling of charged species.[191–193] Thus, it is important to improve the robustness of AFE protocols to enable their reliable application to structure-based drug design problems.

One of the best ways to help us tackle the aforementioned problems is the blinded prediction competitions. They offer a helpful resource to cut down the bias in validation studies and to test practical service of a methodology in a way that more closely resembles CADD in practice.[194] The D3R grand challenges have become one of the most famous blinded competitions. They focus on validating computational methods for modelling of protein-ligand interactions.[194],[195] The Statistical Assessment of Modelling of Proteins and Ligands (SAMPL) is a well-established blinded competition for free energy science in drug discovery.[196] The SAMPL challenge was founded in 2007 and usually asks participants to predict physical chemical properties, such as binding affinities for host-guest systems, or hydration free energies of small drug-like molecules.[197],[198] Host-guest systems attract our interest since they provide manageable milestones towards validation of protocols for modelling protein-ligand binding energetics.[199]

A plethora of computational methods have been examined to predict free energies of binding of the host-guest systems ranging from quantum mechanical.[200],[201] to molecular mechanical approaches[202] Monte Carlo (MC) or Molecular Dynamics (MD) simulations are executed to predict the ensemble averages that yield standard binding free energies. Different approximations lead to numerous ways to predict the binding free energies from molecular simulation trajectories e.g finite difference thermodynamic integration (FDTI)[203], free energy perturbations (FEP)[204], or end-states only variants such as Molecular Mechanics/Poisson–Boltzmann Surface Area (MM-PBSA).[205]

The 6th Statistical Assessment of Modelling of Proteins and Ligands (SAMPL6) competition was launched in September 2017. Our group focused on the host-guest leg of this contest, which requested predictions of standard free energies of binding for 27 guests across 3 different hosts. The host molecules consisted of two octa-acids, OA and TEMOA molecules,[206–209] and a cucurbituril ring clip CB8.[210–213] The octa-acid systems (Figure 3.1) are basket shaped where OA contains four flexible propionate side chains bearing two rotatable single bonds each, while TEMOA contains four methyl groups, which alter the shape of the hydrophobic cavity. CB8 is (Figure 3.1) a more flexible host than OA and TEMOA and is a heteroaromatic multicyclic molecule, chemically related to the cucurbiturils, made of methylene bridges containing eight glycoluril units.[212],[213]

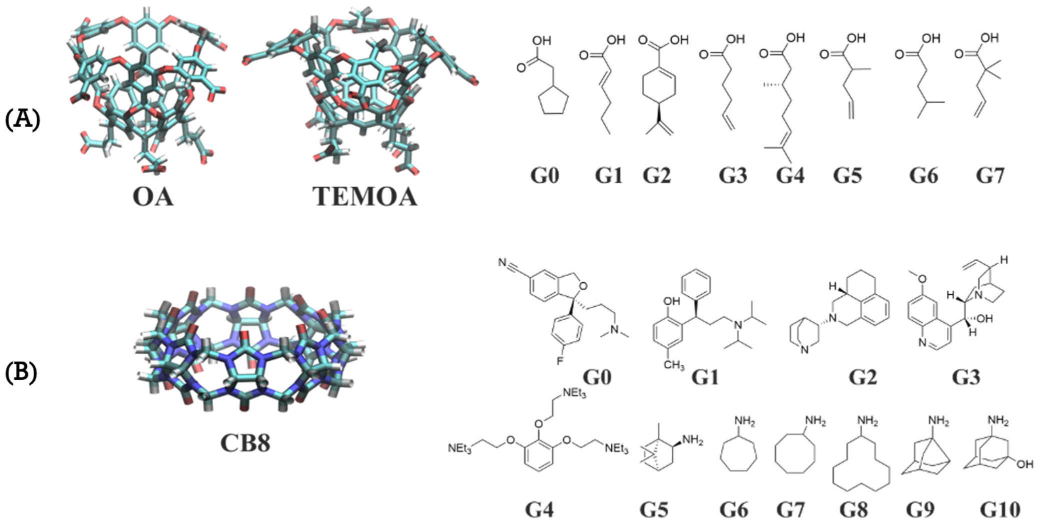


FIGURE 3.1: Depiction of the SAMPL6 host-guest dataset. (A) OA and TEMOA host-guest systems. (B) CB8 host-guest systems.

Additionally, SAMPL6 introduced a SAMPLing challenge focused on evaluating convergence and reproducibility, across codes, of free energy predictions. For this challenge, input files for parameterised host-guests OA-G3, OA-G6 and CB8-G3 were provided and participants were requested to evaluate the convergence of their binding free energy estimates. The three particular guests were chosen because they resemble typical fragments (OA-G3 and OA-G6) and druglike molecules (CB8-G3/Quinine is considered as second-line treatment for malaria).[214] The main goal of this challenge is initially to quantitatively compare the convergence rates of state of the art free energy methods on well-defined host-guest systems and also to evaluate the level of agreement that can be reached by different methods and software starting from identical initial parameters.[215] An overview of the SAMPLing challenge is depicted in Figure 3.2.

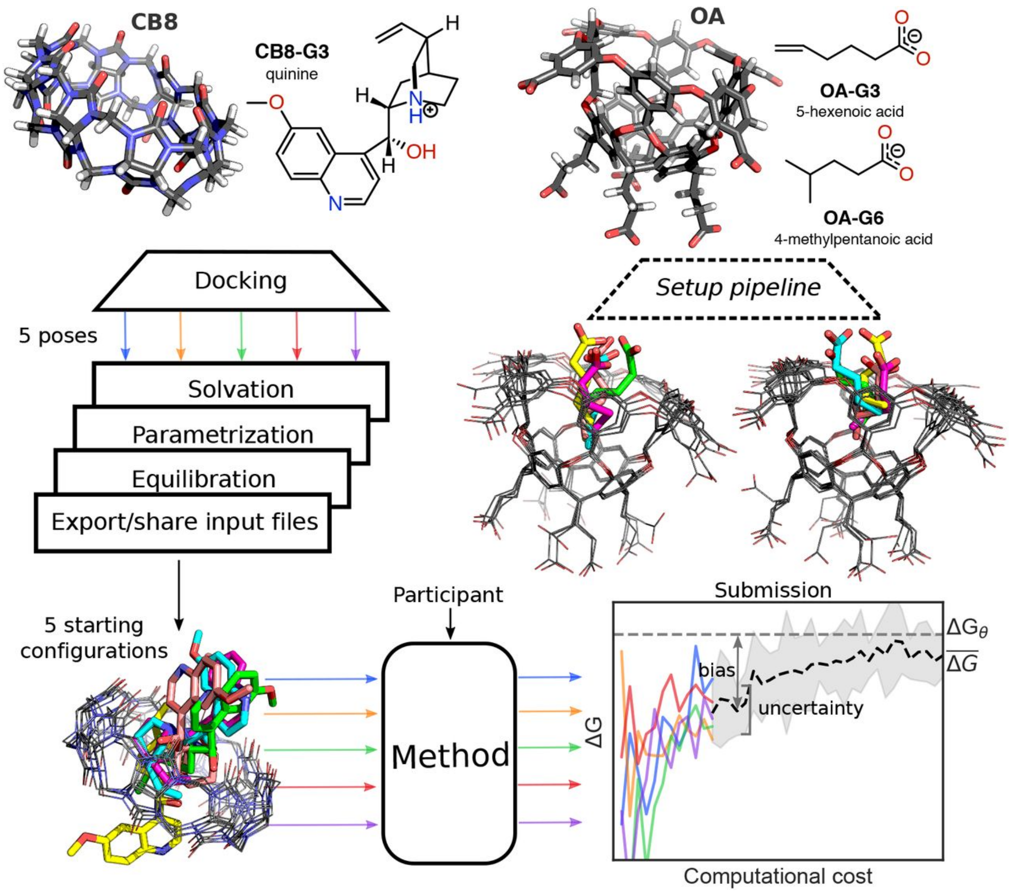


FIGURE 3.2: Overview of the SAMPLIing challenge. The colors used for the 3D structures of the two hosts are grey for carbon atoms, nitrogens in blue, oxygens in red and hydrogens in white. The 2D structures of the guest molecules and the 3D structures of the hosts are illustrated in the protonation state used for the computational predictions. Five different initial conformations for the three host-guest complexes were generated through docking followed by a short equilibration with Langevin dynamics. The 3D structures of these conformations are shown from left to right in the figure and the guest's carbon atoms are colored by conformation. These input files were used by each participant to run their methods in five replicates and submit the free energy trajectories as a function of the computational cost. The resulting submissions were analysed in terms of uncertainty of the mean binding free energy $\overline{\Delta G}$ and its bias with respect to the asymptotic free energy ΔG_{θ} [215]

This chapter summarizes the performance of our free energy code SOMD against the SAMPL6 host-guest dataset, as well as the lessons learned for continuing efforts to improve the robustness of alchemical free energy methods in CADD.

3.2 Methods

Free energy changes were evaluated by means of a double annihilation technique using MD simulations.[53, 54, 182] Figure 3.3 illustrates how this approach is used to evaluate ΔG°_{bind} using a thermodynamic cycle. In the first step (so called *'discharging'* step) the charges of the guest's atoms are turned off both in the solvated phase and in the bound phase, providing the *discharging* free energy changes ΔG^{solv}_{elec} and ΔG^{host}_{elec} respectively. In the second step (so called *'vanishing'* step) a "non-interacting" guest is obtained by switching off the van der Waals parameters of the discharged guest both in solvent and complex phase, giving the *vanishing* free energy changes, ΔG^{solv}_{vdW} and ΔG^{host}_{vdW}, respectively. To prevent the ligand from drifting away from the host cavity, a series of a flat-bottom distance restraints are defined between the guest atom j that is closest to the center of mass of the guest and four host atoms i. The restraint potential is given by Equation 3.1:

$$U^{restr}_{(d_{j1},\, ...,\, d_{jN_{host}})} = \sum_{i=1}^{N_{host}} \begin{cases} 0 & if \ \left|d_{ji} - R_{ji}\right| \leq D_{ji} \\ \kappa_{ij} \left(\left|d_{ji} - R_{ji}\right| - D_{ji}\right)^2 & if \ \left|d_{ji} - R_{ji}\right| > D_{ji} \end{cases}, \quad (3.1)$$

where $U^{restr}_{(d_{j1},\, ...,\, d_{jN_{host}})}$ is the potential energy of the restraint term as a function of the distance between a guest atom j and a set of host atoms i, d_{ji} is the distance between a guest atom j and a host atom i, R_{ji} is the reference distance between host and guest atom, D_{ji} is the restraint deviation tolerance (how much the reference distance can deviate from its original value), k_{ji} is the restraint force constant and N_{host} is the number of host atoms that contribute to the restraint.

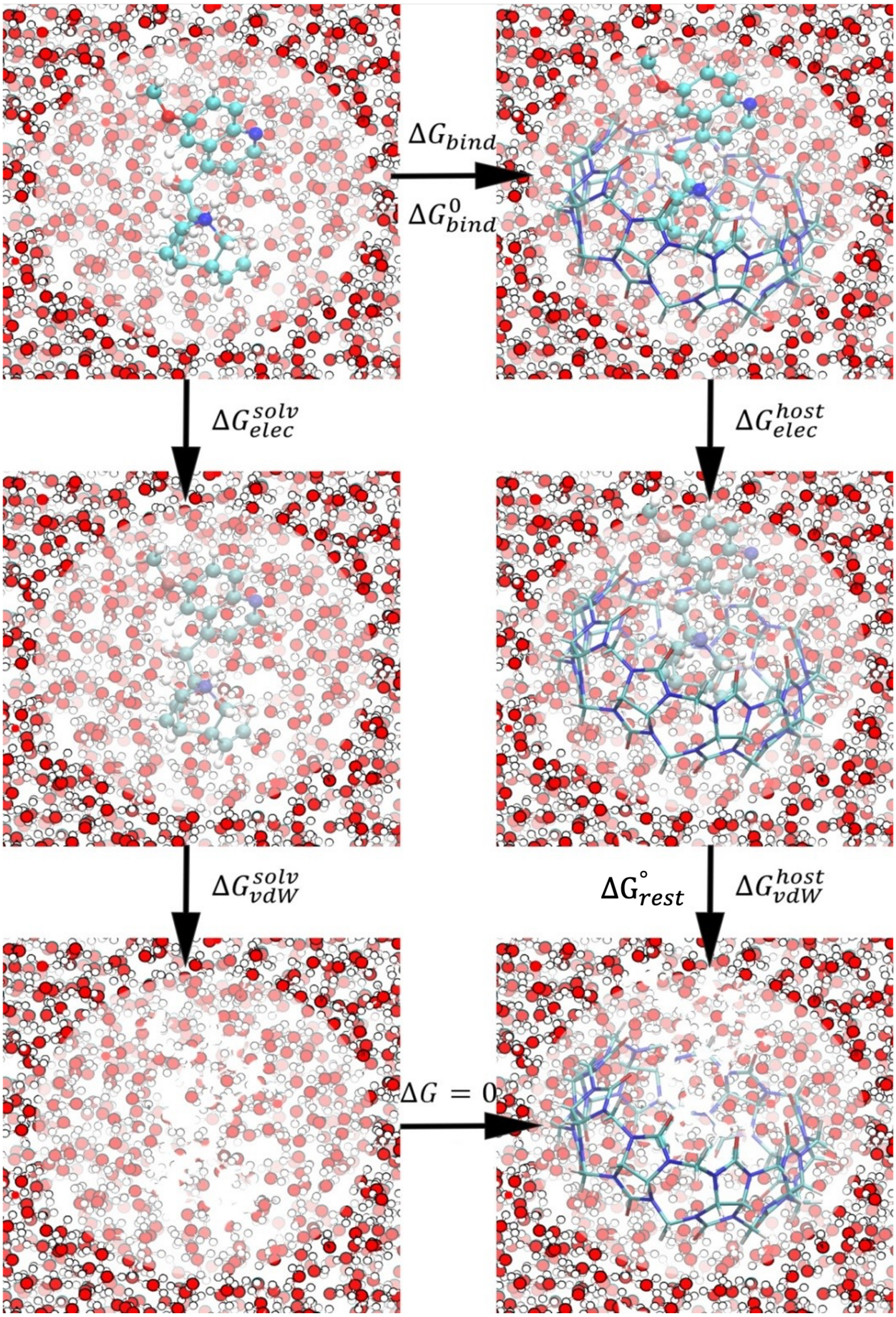


FIGURE 3.3: Thermodynamic cycle for standard binding free energy calculations. Firstly, the fully interacting guest is simulated in a free phase (top left) and a bound phase (top right), then the charges and the van der Waals terms are switched off, resulting in a non-interacting guest in water (bottom left), and bound to the host (bottom right).

From the closure of the thermodynamic cycle (Figure 3.3) the binding free energy ΔG_{bind} is given by Equation 3.2. The free energies of binding computed with Equation 3.2 will be referred to as *Model A* binding energies:

$$\Delta G_{bind}^{ModelA} = \left(\Delta G_{elec}^{solv} + \Delta G_{vdW}^{solv}\right) - \left(\Delta G_{elec}^{host} + \Delta G_{vdW}^{host}\right) . \tag{3.2}$$

Model A does not take into account the contribution of long range dispersions interactions due to the use of non-bonded cutoffs. Thus, to improve over *Model A*, a long-range dispersion correction term is added to the free energy of binding from the simulation trajectories[216]:

$$\Delta G_{bind}^{ModelB} = \Delta G_{bind}^{ModelA} + \left(\Delta G_{LJLRC}^{host} - \Delta G_{LJLRC}^{solv}\right) . \tag{3.3}$$

The Lennard Jones dispersion correction term (LJLRC) can be calculated from the Zwanzig relationship[49]:

$$\mathit{\Delta} G_{\mathrm{LJLRC}}^{\mathrm{X}} = -k_B T \ln\langle \exp[-\beta(U_{\mathrm{LJ,long}}(\mathbf{r}) - U_{\mathrm{LJ,sim}}(\mathbf{r})]\rangle_{\mathrm{X}} + U_{\mathrm{LJ,ana}} , \tag{3.4}$$

where X is the host or solvent, and $U_{LJ,long}$ is the Lennard Jones energy calculated in a post processing step of the *'vanishing'* trajectories generated at λ = 0 and λ = 1, by expanding the range of the typical Lennard Jones cutoff radius in the simulation from 12Å to cover almost the entire box. To define this cutoff, the minimum box length in all directions in the input coordinates is computed and the new cutoff radius is set to $r_{c,long}$ *= 0.95min(Lx, Ly, Lz)/2* to allow for some variations in box size. This permits an averaging of the additional contribution of the long-range potential over the whole trajectory, $U_{LJ,long}$, with respect to $U_{LJ,sim}$ (simulated Lennard Jones term). Additionally an analytical correction over an infinite size box, $U_{LJ,ana}$ is introduced, which is given by Equation 3.5:

$$U_{\mathrm{LJ,ana}} = 8\pi\rho \sum_{i}^{N_{\mathrm{solute}}} \sum_{j}^{N_{\mathrm{solv}}} \left[\frac{\epsilon_{ij}\sigma_{ij}^{12}}{9r_{\mathrm{c}}^{9}} - \frac{\epsilon_{ij}\sigma_{ij}^{6}}{3r_{\mathrm{c}}^{3}} \right] , \tag{3.5}$$

where p is the solvent density in mol Å^{-3}, N_{solute} is the total number of guest atoms, N_{solv} the number of solvent molecules, ϵ_{ij} is the Lennard Jones well depth, expressed in kcal mol^{-1} and σ_{ij} is the Lennard Jones distance in Å computed with the Lorentz-Berthelot combining rule[38].

Additionally, a free energy correction term is introduced to relate the volume available to the restrained but non-interacting ligand to standard state conditions. This leads to Equation 3.6 for predictions of binding free energies via *Model B*:

$$\Delta G_{bind}^{\circ,ModelB} = \Delta G_{bind}^{ModelA} + \left(\Delta G_{LJLRC}^{host} - \Delta G_{LJLRC}^{solv}\right) + \Delta G_{restr}^{0} \ , \tag{3.6}$$

where ΔG_{restr}° is the free energy cost for imposing the host-guest restraint which is given by Equation 3.7:

$$\Delta G_{restr}^{\circ} = \ -k_B T \ln \left(\frac{Z_{HG_{ideal}}}{Z_{H,solv}\, Z_{G,gas}} \right) \ , \tag{3.7}$$

where $Z_{HG_{ideal}}$ is the configuration integral for the restrained decoupled guest bound to the host, $Z_{H,solv}$ is the configuration integral for the solvated host and $Z_{G,gas}$ is the configuration integral for the guest in an ideal thermodynamic state. If it is assumed that the restraint potential is decoupled from the solvent and host degrees of freedom, then the equation 3.7 can be simplified to:

$$\Delta G_{\mathrm{restr}}^{\circ} = -k_{\mathrm{B}} T \ln \left(\frac{Z_{\mathrm{G_{ideal}},\mathrm{solv}}}{Z_{\mathrm{G,gas}}} \right) \ , \tag{3.8}$$

where $Z_{\mathrm{G_{ideal}},solv}$ is the configuration integral for the decoupled guest. As the intermolecular interactions are absent from the guest in the thermodynamic states described in Equation 3.7 and the restraint do not prevent rotational motions, the internal and rotational contributions to the configuration integrals are cancelled out and the only term left is the translational contribution to the configuration integral. A standard volume of measurement V° is used for $Z_{\mathrm{G,gas}}$, where the 1M dilute solute convention correlates to V° = *1660 Å^{-3} mol^{-1}*. Thus Equation 3.8 becomes:

$$\Delta G_{\mathrm{restr}}^{\circ} = -k_{\mathrm{B}} T \ln \left(\frac{V^{\mathrm{restr}}}{V^{\circ}} \right) \ , \tag{3.9}$$

where V^{restr} can be computed by numerically integrating Equation 3.9:

$$V^{\text{restr}} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} dx_j \, dy_j \, dz_j \exp(-\beta U^{restr}(d_{j1}, \ldots, d_{jN_{host}})) \,. \quad (3.10)$$

Finally, *Model C* was constructed by creating an empirical correction term to account for systematic errors due to finite size artefacts and inaccuracies in potential energy functions. Linear regression models were obtained by correlating past SAMPL5 binding free energies computed with SOMD to experimental data, leading to equation 3.11 that computes *Model C* binding free energies:

$$\Delta G^{\circ,ModelC}_{bind} = \frac{\Delta G^{\circ,ModelB}_{bind} - \beta}{\alpha} \,, \quad (3.11)$$

where α and β are the slope and intercept of the linear regression model. SAMPL5 featured the same hosts *OA* and *TEMOA* but a different host *CB7*. Thus, separate regression models were determined for use with *OA*, *TEMOA* or *CB8* hosts. The parameters are given in the Appendix.

3.3 Preparation of host-guest input files for free energy calculations

The SAMPL6 organizers provided mol2 files for hosts, *OA*, *TEMOA* and *CB8*, and ligands, depicted in Fig. 3.1. Each file had the same Cartesian frame of reference and docking was performed with OpenEye toolkit[217], [218], [219] to predict the most likely binding mode. Experimental measurements were performed at a pH 11.7 ± 0.1 at 298 K in presence of a buffer of 10 mM Na_3PO_4 for *OA* and *TEMOA*. *CB8* was measured at pH 7.4 ± 0.1 at 298 K with 25 mM Na_3PO_4 buffer. To understand the influence of the buffer on binding free energy predictions, two different sets of input files were prepared, leading to *no-buffer* and *buffer* setups.

In the *no-buffer* simulations, the presence of the additional Na_3PO_4 buffer was neglected. *OA*, *TEMOA* and *CB8* host-guest systems were parametrized starting from the mol2 host and guest's files. *OA* and *TEMOA* molecules force field and charge parameters were retrieved by processing SAMPL5 topology and coordinate files using the python module *parmed*.[220] Thus, tleap[168] was used to create the host-guest complex input files. The combined host-guest complex mol2 file was loaded in tleap along with host force field parameters and GAFF 1.8 for the ligand[170],[171]. The system was solvated in a cubic box with TIP3P water molecules[173], with a minimum distance between the solute and the box of 12 Å. Counter ions were added to neutralize the total net charge. The same approach was followed for parametrising the ligand in the free phase.

Next an equilibration protocol was applied to relax the box size. Initially, energy minimization of the entire system was implemented with 100 steps of steepest descent gradients, using sander[168]. Then, the positions of the solute molecules were restrained with a force constant of 10 kcal mol^{-1} Å^{-2} while water molecules were allowed to equilibrate in an NVT ensemble, 200 ps at 298 K, followed by a NPT equilibration for further 200 ps at 1 atm pressure. Finally, a 2 ns NPT MD simulation was run with the SOMD software (revision 2017.1.0) to reach a final density of about 1 g cm^{-3}[221],[222]. The final coordinate files were retrieved with cpptraj. The edge length of the solvated guest boxes was about 35 Å, whereas the boxes of the complex systems had an edge length of about 50 Å.

For the second set of simulations, additional counter ions were added to mimic the presence of the buffer in the experiments. However, Na_3PO_4 was modelled by NaCl as force-field parameters for multivalent ions were not available. Thus, for *OA* and *TEMOA* systems, the 10 mM sodium phosphate buffer was modelled with 60 mM of NaCl to match the ionic strength of the solution used for the experiments. Starting from the complex phase files, created as described previously, 4 additional Na^+ and 4 Cl^- ions were added to each system, using tleap. The same equilibration protocol was reapplied to adjust the placement of the counter ions. For the preparation the solvated phase, the host molecule was extracted from an equilibrated host-guest box and the host's heavy atoms were replaced with water molecules. After equilibration the final solvated phase system had the same amount of Na^+ and Cl^- ions as in the host-guest complex system, and a similar box size dimension. The same procedure was followed for *CB8*. In this case, 25 mM Na_3PO_4 were matched with 150 mM NaCl, thus 8 Na^+ and 8 Cl^- ions were added to each *CB8* host-guest system.

3.4 SAMPL6 simulation protocols

For the octa-acid hosts, both complex and solvated phase *discharging* steps were run with nine equidistant λ windows. Twelve λ windows (0.00, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.95, 1.00) were employed for the *vanishing* step, both in bound and solvated phase. For the *CB8* host the bound and solvated phase *discharging* steps have been run with nine equidistant λ windows. The solvated *vanishing* step was carried out with the same windows setup as for the octa-acid guests. The bound *vanishing* step was carried out with sixteen λ windows (0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.70, 0.85, 1.00) as preliminary runs indicated a need for greater number of windows to obtain reliable free energy changes.

All the simulations were run for duration of 8 ns with SOMD in an NPT ensemble. Temperature control was achieved with an Andersen Thermostat with a coupling constant of 10 ps^{-1}.[48] Pressure control was maintained by a Monte Carlo barostat that attempted isotropic box edge scaling every 100 fs. A 12 Å atom-based cutoff distance for the non-bonded interactions was used, using a Barker Watts reaction field with dielectric constant of 78.3.[223] In the bound phase the restraints parameters of eq. 3.1 were: R_{ji} = 5 Å, D_{ji} = 2 Å and k_{ji}= 10 kcal mol^{-1} $Å^{-2}$ for all the octa-acid systems, while R_{ji} = 7 Å, D_{ji} = 2 Å and k_{ji} = 10 kcal mol^{-1} $Å^{-2}$ were chosen for the *CB8* simulations.

3.5 SAMPLing simulation protocols

For the SAMPLing leg of the challenge topologies and coordinate file for five replicates of OA-G3, OA-G6 and CB8-G3 were provided from the organizers for both the complex phase and the solvated phase simulations. All simulations were run for 20 ns per window using SOMD with simulation parameters identical to those used for SAMPL6 unless otherwise mentioned.

3.6 Estimation of free energy of binding and evaluation of dataset metrics

Free energy changes were computed using the multistate Bennet acceptance ratio (MBAR) method.[50] To achieve a more robust estimation of free energies, each simulation was repeated multiple times, using different initial velocities drawn from the Maxwell-Boltzmann distribution. Unless otherwise mentioned, the reported binding free energies are the mean of three runs, and statistical uncertainties are given one standard error of the mean as shown in Equation 3.12:

$$err(\Delta G) = \frac{\sigma}{n}. \tag{3.12}$$

As described in Bosisio *et al*[182] for each model a population distribution for the determination coefficient R^2, the mean unsigned error MUE and the Kendall τ parameters were computed by bootstrapping each free energy predictions for each host-guest dataset ten thousand times. The determination coefficient R^2 is the fraction of the variance of the dependent variable that can be predicted from the independent variable. For a dependent variable y and an independent variable x, R^2 can be expressed through Equation 3.13:

$$R^2 = \left(\frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2]) - [n\sum y^2 - (\sum y)^2]}}\right). \tag{3.13}$$

The MUE is a measure of difference between two continuous variables, in our case the predicted and the experimental binding free energy. It is given by Equation 3.14 for a set of n datapoints:

$$MUE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} . \tag{3.14}$$

Kendall τ is a measure of the degree of similarity between two quantities, here the predicted and the experimental binding free energy. Kendall τ will be high when observations have similar relative positions. Kendall τ is dependent on the type of the observation pairs, that can be concordant or discordant. Let's assume that *(x_1,y_1), (x_2,y_2), . . . , (x_n,y_n)* are the set of observations of the random variables X and Y respectively, where all the values of x_i and y_i are unique. The pair of observations *(x_i,y_i)* and *(x_j,y_j)*, where $i <$ j, is called concordant if the sort order by x and y agree, i.e. if both x_i>x_j and y_i>y_j or vice versa. If x_i<x_j and y_i>y_j or vice versa the pair of observations is called discordant. Finally, if x_i=x_j and y_i=y_j the pair is neither concordant nor discordant. The Kendall τ coefficient is depicted in Equation 3.15 as:

$$\tau = \frac{C - D}{n(n-1)\backslash 2} , \tag{3.15}$$

where C is the number of concordant pairs and D is the number of disconcordant pairs. The resulting distributions may not be symmetric around the mean, thus uncertainties are reported with a 95 % confidence interval. Additionally, for the SAMPLing leg of the challenge, binding free energies were evaluated using *Model B* by skipping the first 1.5 ns of each window, and using 1 to 100% of the rest of the dataset. Uncertainties were taken as the standard deviation output from pymbar and were propagated to obtain an uncertainty for the reported standard free energy of binding. The total wall-clock time was also estimated by summing up the wall-clock time for each λ window, in each phase and simulated process. The number of iterations was retrieved as the sum of the number of time-steps for each simulated process. For each host-guest replica, 459,995,400 energy evaluations were carried out with an average wall-clock time of 245 hours for *CB8* systems and 190 hours for *OA*. All input files for the SAMPL6 and SAMPLing protocols are publically available in the repository `https://github.com/michellab/SAMPL6inputs`.

3.7 SAMPL6 results

Results for the full SAMPL6 dataset are shown in Figure 3.4 for each model without and with a buffer setup. As judged by MUE, *Model A/no-buffer* is the least accurate protocol, with a MUE value ca. 5.7 kcal mol^{-1}. *Model A/buffer* offers some improvement in accuracy, with the MUE decreasing to ca. 5.1 kcal mol^{-1}. Addition of long-range dispersions and standard state correction terms in *Model B* decreases MUE further (MUE ca. 3.9 and 3.4 kcal mol^{-1} for the *no-buffer* and *buffer* setups respectively). *Model C* improves over *Model B* with MUE values ca. 1.4 and 1.6 kcal mol^{-1} for the *no-buffer* and *buffer* setups respectively. Thus, the additional counter-ions in the *buffer* setup improve the accuracy for *Model A* and *Model B* but not for *Model C*. This could be because the SAMPL5 calculations were carried out with a *no-buffer* setup,[182] and the empirical correction terms used in *Model C* do not transfer to a *buffer* setup.

Ranking of the protocols according to correlation with experimental data produces a different result. *Model A/no-buffer* and *Model B/no-buffer* perform similarly well with R^2 and τ values ca. 0.6. A small drop in predictive power is observed for *Model C/no-buffer* but this is only significant for R^2. This decrease is observed because the empirical correction term helps to bring the *OA* host-guest binding energies in line with the experimental values, but leads to a trend of underestimating the *CB8* binding energies. The use of *buffer* also appears deteriorates the predictive power, with all *buffer* protocols giving significant drops in R^2 and τ parameters with respect to the equivalent *no-buffer* protocol.

Inspection of the results for the *OA* subset (Table 3.2 and Table 3.3) shows that *Model B* and *Model C* significantly improve the MUE over *Model A* but not for R^2 and τ metrics that are ca. 0.7 and 0.5 respectively. The *buffer* protocol worsens MUE over the *no-buffer* protocol but does not affect the predictive power. The same picture holds for the*TEMOA* subset, with improvements for MUE observed as correction terms are introduced. In addition, switching from *no-buffer* to *buffer* deteriorates the MUE for *Model A* and *Model B*. The R^2 and τ metrics are higher than *OA* (ca. 0.9 and 0.8) and insensitive to the various protocols. For the *CB8* subset, dramatic improvements in MUE are also observed upon switching from *Model A* to *Model B* and *Model C* (*Model A/no-buffer* MUE ca. 7.3 kcal mol^{-1} vs *Model C/no-buffer* MUE ca. 1.6 kcal mol^{-1}). Unlike for the *OA* and *TEMOA* hosts, switching from a *no-buffer* to *buffer* setup significantly improves the MUE for *Model A* and *Model B*, but not for *Model C* where the MUE gets worse. Thus, the *buffer* effects depend on the nature of the host-guest systems. For the octa-acid guests, the guests are negatively charged acids and explicit modelling of a buffer favors the binding process (average change in binding energies of -0.9 kcal mol^{-1} for *Model B*). For the cucurbit-uril host, the guests are positively charged amines and explicit modelling of a buffer disfavors the binding process (average change in binding energies of +3.1 kcal mol^{-1} for *Model B*). The effect is particularly pronounced for some *CB8* guests, e.g. the binding energies of G3, G4 and G7 increase by more than 4 kcal mol^{-1} upon switching from a *no-buffer* to *buffer* protocol. None of the models tested yield significant predictive power with R^2 and τ metrics ca. 0.1.

The largest outliers for *CB8* are guests G3, G4, G5 and G8. In particular, the free energies of binding of G3, G5 and G8 are lower than the experimental data by about 10 kcal mol^{-1} with *Model A/no-buffer* or *Model B/no-buffer*. The statistical errors are also bigger than for the *OA* and *TEMOA*, suggesting greater challenges for converging free energy changes in *CB8* over the simulated time-scales. However, using a *buffer* protocol lowers free energies of binding, and by up to ca. 5 kcal mol^{-1} for G3 and G8.

Among octa-acids the models correctly capture interesting tendencies in the ITC measurements. For instance, the models correctly predict that G7 binds significantly worse to *TEMOA* than to *OA*. The bulkiness of the two β methyl groups to the carboxylic acid moiety blocks positioning of the guest in the smaller *TEMOA* cavity (Fig 1A). The most significant outlier is G2 for which the models are not able to reproduce the significantly lowered binding energetics for *TEMOA* vs *OA*. A possible reason for this inconsistency is that the different ring puckering motions of the cyclohexenil moiety in G2 may have been poorly sampled with the simulation protocols used.

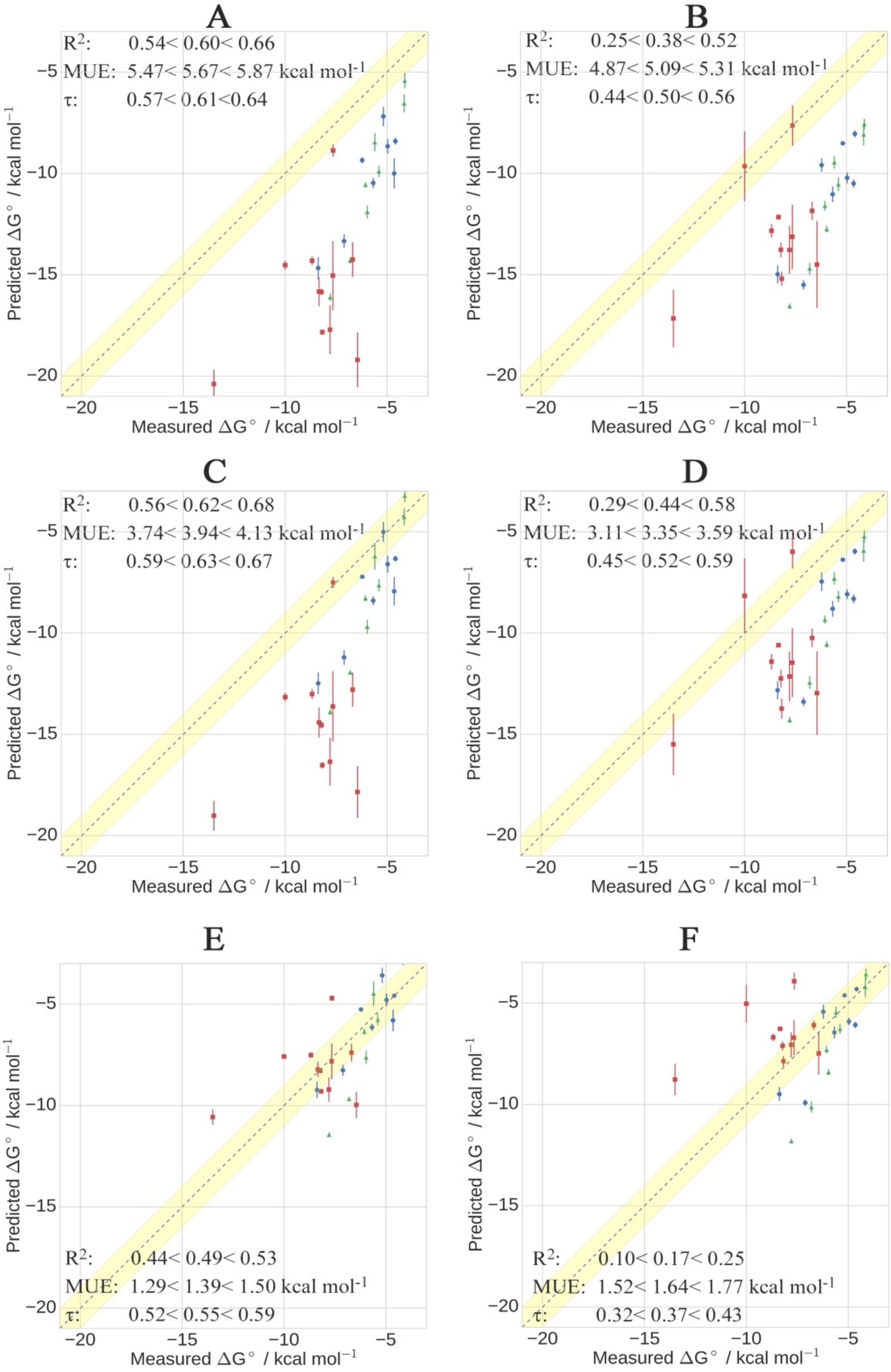


FIGURE 3.4: Comparison of the predicted and measured binding free energies for A) *Model A/no-buffer* B) *Model A/buffer* C) *Model B/no-buffer* D) *Model B/buffer* E) *Model C/no-buffer* F) *Model C/buffer* for the 27 host-guest systems. The grey line denotes perfect correlation between predictions and measurements, while the yellow shaded region indicates a ±1 kcal mol^{-1} error bound. *OA* systems are illustrated in blue circles, *TEMOA* in green triangles and *CB8* in red squares.

OA				
Guest	$\Delta G_{bindModelA}$	$\Delta G_{bindModelB}$	$\Delta G_{bindModelC}$	$\Delta G_{bindExperimental}$
G0	-10.5 ± 0.3	-8.4 ± 0.2	-6.1 ± 0.2	-5.7 ± 0.1
G1	-10.0 ± 0.7	-7.9 ± 0.7	-5.8 ± 0.5	-4.7 ± 0.1
G2	-14.7 ± 0.5	-12.5 ± 0.5	-9.2 ± 0.4	-8.4 ± 0.1
G3	-7.2 ± 0.5	-5.0 ± 0.5	-3.6 ± 0.4	-5.2 ± 0.1
G4	-13.3 ± 0.3	-11.2 ± 0.4	-8.3 ± 0.3	-7.1 ± 0.1
G5	-8.4 ± 0.2	-6.3 ± 0.2	-4.6 ± 0.1	-4.6 ± 0.1
G6	-8.7 ± 0.4	-6.6 ± 0.4	-4.8 ± 0.3	-5.0 ± 0.1
G7	-9.4 ± 0.2	-7.2 ± 0.1	-5.3 ± 0.1	-6.2 ± 0.1
R^2	$0.62 < 0.75 < 0.85$	$0.62 < 0.73 < 0.84$	$0.62 < 0.74 < 0.84$	
MUE	$4.17 < 4.41 < 4.66$	$2.15 < 2.37 < 2.60$	$0.65 < 0.82 < 1.00$	
τ	$0.43 < 0.54 < 0.64$	$0.43 < 0.54 < 0.64$	$0.43 < 0.54 < 0.64$	
TEMOA				
G0	-10.6 ± 0.1	-8.3 ± 0.2	-6.4 ± 0.2	-6.1 ± 0.1
G1	-11.9 ± 0.3	-9.7 ± 0.4	-7.6 ± 0.3	-6.0 ± 0.1
G2	-14.3 ± 0.1	-11.9 ± 0.1	-9.7 ± 0.1	-6.8 ± 0.1
G3	-8.5 ± 0.5	-6.2 ± 0.7	-4.5 ± 0.2	-5.6 ± 0.1
G4	-16.1 ± 0.2	-13.9 ± 0.1	-11.4 ± 0.1	-7.8 ± 0.1
G5	-6.5 ± 0.4	-4.3 ± 0.4	-2.7 ± 0.4	-4.2 ± 0.1
G6	-9.9 ± 0.3	-7.6 ± 0.3	-5.8 ± 0.3	-5.4 ± 0.1
G7	-5.4 ± 0.4	-3.2 ± 0.3	-1.8 ± 0.3	-4.1 ± 0.1
R^2	$0.90 < 0.93 < 0.96$	$0.89 < 0.93 < 0.97$	$0.91 < 0.94 < 0.96$	
MUE	$4.47 < 4.66 < 4.84$	$2.50 < 2.67 < 2.84$	$1.58 < 1.72 < 1.86$	
τ	$0.85 < 0.86 < 0.87$	$0.78 < 0.85 < 0.86$	$0.85 < 0.86 < 0.87$	
CB8				
G0	-14.3 ± 0.9	-12.8 ± 0.9	-7.4 ± 0.4	-6.7 ± 0.1
G1	-8.9 ± 0.3	-7.5 ± 0.3	-4.7 ± 0.1	-7.7 ± 0.1
G2	-15.0 ± 1.7	-13.6 ± 1.7	-7.8 ± 0.9	-7.7 ± 0.1
G3	-19.2 ± 1.4	-17.8 ± 1.3	-10.0 ± 0.7	-6.5 ± 0.1
G4	-17.7 ± 1.2	-16.4 ± 1.2	-9.2 ± 0.6	-7.8 ± 0.1
G5	-17.8 ± 0.1	-16.5 ± 0.2	-9.3 ± 0.1	-8.2 ± 0.1
G6	-15.8 ± 0.7	-14.4 ± 0.7	-8.2 ± 0.4	-8.3 ± 0.1
G7	-14.5 ± 0.2	-13.2 ± 0.2	-7.6 ± 0.1	-10.0 ± 0.1
G8	-20.4 ± 0.7	-19.0 ± 0.7	-10.6 ± 0.4	-13.5 ± 0.1
G9	-14.3 ± 0.2	-13.0 ± 0.3	-7.5 ± 0.1	-8.7 ± 0.1
G10	-15.9 ± 0.2	-14.5 ± 0.2	-8.3 ± 0.1	-8.2 ± 0.1
R^2	$0.04 < 0.12 < 0.23$	$0.04 < 0.12 < 0.23$	$0.04 < 0.12 < 0.23$	
MUE	$6.90 < 7.33 < 7.76$	$5.57 < 5.99 < 6.42$	$1.40 < 1.57 < 1.76$	
τ	$-0.09 < 0.04 < 0.20$	$-0.09 < 0.05 < 0.20$	$-0.09 < 0.05 < 0.20$	

TABLE 3.1: Results for all three models (*no-buffer* protocol) for individual host-guest families. Energies and MUE are reported in kcal/mol

OA				
Guest	$\Delta G_{bindModelA}$	$\Delta G_{bindModelB}$	$\Delta G_{bindModelC}$	$\Delta G_{bindExperimental}$
G0	-11.0 ± 0.4	-8.8 ± 0.4	-6.5 ± 0.3	-5.7 ± 0.1
G1	-10.0 ± 0.2	-8.3 ± 0.2	-6.1 ± 0.2	-4.7 ± 0.1
G2	-15.0 ± 0.4	-12.8 ± 0.5	-9.5 ± 0.3	-8.4 ± 0.1
G3	-8.5 ± 0.1	-6.4 ± 0.1	-4.6 ± 0.0	-5.2 ± 0.1
G4	-15.5 ± 0.2	-13.4 ± 0.2	-9.9 ± 0.2	-7.1 ± 0.1
G5	-8.1 ± 0.2	-6.0 ± 0.2	-4.3 ± 0.1	-4.6 ± 0.1
G6	-10.2 ± 0.3	-8.1 ± 0.2	-5.9 ± 0.2	-5.0 ± 0.1
G7	-9.6 ± 0.3	-7.5 ± 0.5	-5.4 ± 0.4	-6.2 ± 0.1
R^2	0.67 < 0.74 < 0.80	0.63 < 0.71 < 0.78	0.63 < 0.71 < 0.78	
MUE	4.97 < 5.14 < 5.31	2.87 < 3.05 < 3.23	0.95 < 1.08 < 1.22	
τ	0.50 < 0.56 < 0.64	0.43 < 0.52 < 0.64	0.43 < 0.52 < 0.64	
TEMOA				
G0	-11.6 ± 0.2	-9.3 ± 0.2	-7.3 ± 0.2	-6.1 ± 0.1
G1	-12.8 ± 0.2	-10.6 ± 0.2	-8.4 ± 0.2	-6.0 ± 0.1
G2	-14.7 ± 0.3	-12.4 ± 0.3	-10.1 ± 0.3	-6.8 ± 0.1
G3	-9.5 ± 0.3	-7.3 ± 0.3	-5.5 ± 0.3	-5.6 ± 0.1
G4	-16.6 ± 0.1	-14.3 ± 0.1	-11.8 ± 0.1	-7.8 ± 0.1
G5	-8.1 ± 0.5	-5.9 ± 0.5	-4.2 ± 0.5	-4.2 ± 0.1
G6	-10.5 ± 0.3	-8.2 ± 0.3	-6.3 ± 0.3	-5.4 ± 0.1
G7	-7.6 ± 0.3	-5.2 ± 0.3	-3.6 ± 0.3	-4.1 ± 0.1
R^2	0.89 < 0.92 < 0.95	0.89 < 0.93 < 0.96	0.89 < 0.93 < 0.96	
MUE	5.48 < 5.66 < 5.84	3.23 < 3.41 < 3.59	1.51 < 1.63 < 1.77	
τ	0.78 < 0.84 < 0.85	0.78 < 0.85 < 0.86	0.78 < 0.84 < 0.85	
CB8				
G0	-11.9 ± 0.5	-10.2 ± 0.5	-6.1 ± 0.2	-6.7 ± 0.1
G1	-7.6 ± 1.0	-6.0 ± 0.8	-3.9 ± 0.4	-7.7 ± 0.1
G2	-13.1 ± 1.6	-11.5 ± 1.7	-6.7 ± 0.9	-7.7 ± 0.1
G3	-14.5 ± 2.2	-13.0 ± 2.1	-7.5 ± 1.1	-6.5 ± 0.1
G4	-13.8 ± 1.2	-12.1 ± 1.2	-7.1 ± 0.6	-7.8 ± 0.1
G5	-15.2 ± 0.4	-13.7 ± 0.5	-7.9 ± 0.3	-8.2 ± 0.1
G6	-12.2 ± 0.1	-10.6 ± 0.0	-6.3 ± 0.0	-8.3 ± 0.1
G7	-9.7 ± 1.7	-8.2 ± 1.8	-5.0 ± 0.9	-10.0 ± 0.1
G8	-17.2 ± 1.4	-15.5 ± 1.5	-8.8 ± 0.8	-13.5 ± 0.1
G9	-12.3 ± 0.3	-11.4 ± 0.4	-6.7 ± 0.2	-8.7 ± 0.1
G10	-13.8 ± 0.4	-12.2 ± 0.4	-7.1 ± 0.2	-8.2 ± 0.1
R^2	0.00 < 0.13 < 0.33	0.00 < 0.13 < 0.35	0.00 < 0.13 < 0.35	
MUE	4.07 < 4.59 < 5.13	2.96 < 3.52 < 4.08	1.78 < 2.06 < 2.34	
τ	-0.12 < 0.06 < 0.27	-0.09 < 0.10 < 0.30	-0.09 < 0.10 < 0.30	

TABLE 3.2: Results for all three models (*buffer* protocol) for individual host-guest families. Energies and MUE are reported in kcal/mol

3.8 SAMPLing results

Seven different alchemical or physical binding free energy methodologies were implemented using OpenMM[222], NAMD[224], GROMACS[225] and AMBER.[168] Three submissions used weighted ensemble (OpenMM/REVO), alchemical nonequilibrium switching (GROMACS/NS-DS/SB) or potential of mean force (AMBER/APR). The other four (OpenMM/XREX, GROMACS/EE, NAMD/BAR, and our submission) are based on the double decoupling methodology and only differ in the enhancing sampling strategies and protocols that were implemented. Detailed information about each protocol can be found in Rizzi *et al.*[226] The force-fields and the charges used were identical for all the calculations, but there were small differences between our model and the other models. The main differences were in terms of the treatment of long range interactions as SOMD does not support particle-mesh Ewald (PME) method used by the other participants. In addition, our model constrained all the bonds to their equilibrium value rather than constraining only the bonds involving hydrogens. We are also using a Lennard-Jones cutoff of 12 Å instead of 10 Å. All of the standard binding free energies were computed with respect to a standard concentration of 1M.

Figure 3.5 illustrates the overall results of the seven different methodologies for the three host-guest systems. The five replicate calculations for the different conformations of each method are always within 0.1-0.6 kcal/-mol for OA-G6 and within 0.1-0.4kcal/mol for OA-G3 with the exception of OpenMM/REVO and this level of convergence was achieved in less than 400x10^6 force evaluations. However, for CB8-G3 the agreement between replicates of the same method is slightly worse. This suggests that if reasonable computational resources are available, absolute binding free energy calculations can achieve convergence for this class of systems.

It is also interesting to compare the predictions from the five independent runs with the ITC measurements from the host-guest challenge.[227–229] The corresponding experimental data yielded binding free energies of -5.18 ± 0.02 kcal/mol for OA-G3, -4.97 ± 0.02 kcal/mol for OA-G6 and -6.45 ± 0.06 kcal/mol for CB8-G3. The computational results were more negative on average by 1.2 kcal/mol for OA-G3, -2.1 kcal/mol OA-G6 and -4.4 kcal/-mol CB8-G3 respectively. These observations were in line with the SAMPL6 host-guest challenge. However, it should be noted that the ionic stengths of SAMPLing systems were higher than in experimental conditions (60 mM for OA-G3 and OA-G6 versus 41.25 mM and 150 mM for CB8-G3 versus 57.8 mM) and as we have shown in the SAMPL6 results this could make a difference for the two octa-acid systems but not for CB8-G3 in terms of accuracy.

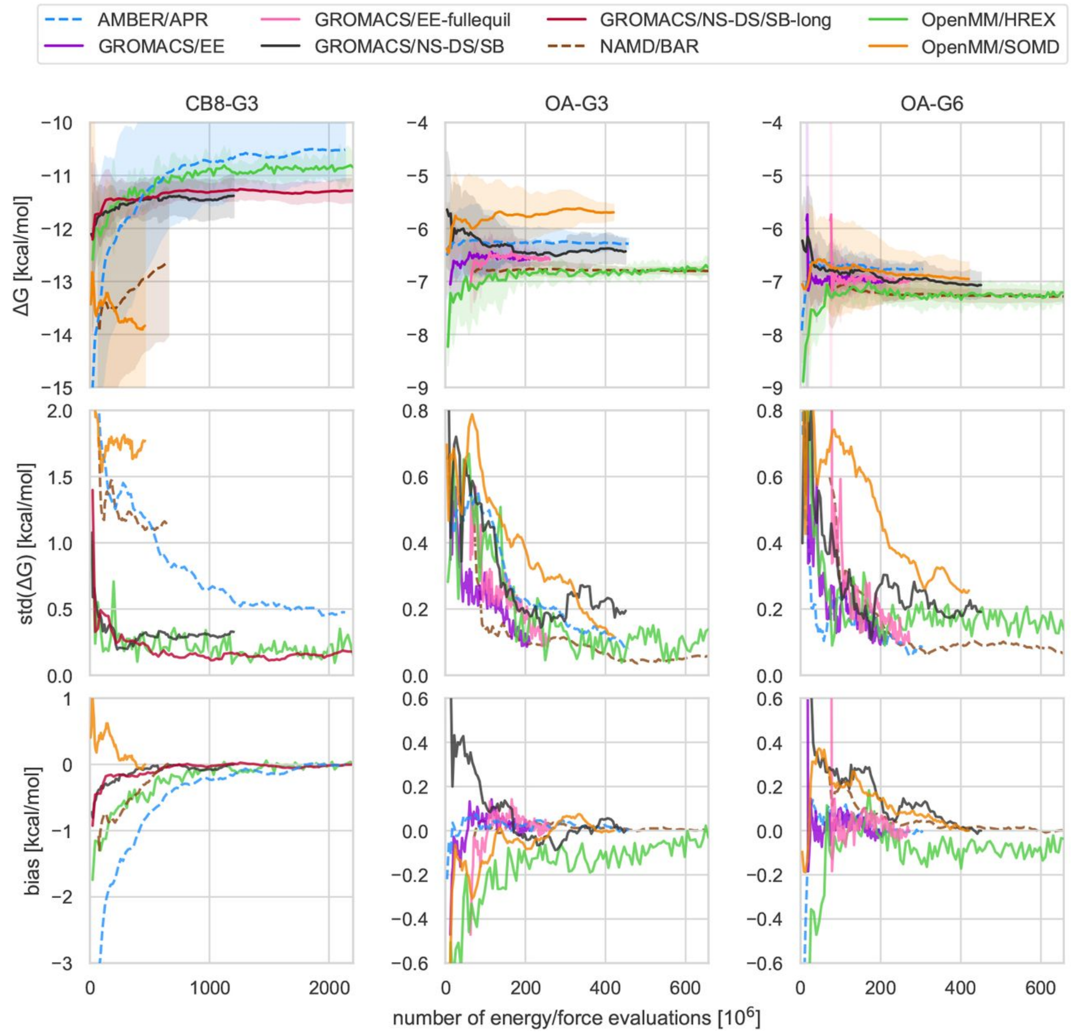


FIGURE 3.5: Mean free energy, standard deviation and bias as a function of computational effort. The mean free energies of binding and 95% t-based confidence intervals computed from the five replicates of CB8-C3 (left), OA-G3 (center) and OA-G6 (right) for all submissions are represented by the trajectories and the shaded areas in the top row. The standard deviation and bias as a function of the computational cost are illustrated in the second and third rows respectively. Given the differences in the simulation parameters between different approaches, the finite-time bias is estimated by assuming that the theoretical binding free energy of the calculation is the final value of its mean free energy. Thus, the bias can either go to zero, or it can be underestimated if the simulation is not converged. Adapted from Rizzi *et al*.[215]

Convergence plots for the calculated binding free energies of the three host guests CB8-G3, OA-G3 and OA-G6 compared to the ones computed by the organizers using the software YANK are presented in Figure 3.6. Figure 3.6A shows that for CB8-G3 the binding free energy obtained using the full simulation dataset is -13.8 ± 0.7 kcal mol^{-1}. Although the uncertainties are high, the mean free energy quickly settles around -14 kcal mol^{-1} and similar estimates would have been obtained with about 20% of the simulation duration. The calculated binding free energies are consistent with those obtained for the host-guest part of the SAMPL6 challenge (-13.0 ± 2.1 kcal mol^{-1}, 3.2). In general, our final prediction is far more negative than the other participants with the exception of OpenMM/REVO and NAMD/BAR and from the experimental result (-6.5 ± 0.1 kcal mol^{-1}). For instance, the SAMPLing reference binding free energy computed by the organizers using the software YANK is significantly different and more precise (-10.8 ± 0.2 kcal mol^{-1}). The reference value is also in better agreement with the experimental data, though considerable differences remain. It appears that at least 60% of the simulation duration is needed to eliminate drifts in the running average for the reference calculation.

For OA-G3 (Figure 3.6B) the binding free energies computed with SOMD and by the other methods are similarly precise. For instance, our prediction compared to the SAMPLing reference converge to -5.7 ± 0.1 kcal mol^{-1} and -6.7 ± 0.1 kcal mol^{-1} respectively. The SOMD SAMPLing free energies are as precise but more accurate than the SOMD SAMPL6 free energies (-6.4 ± 0.1 kcal mol^{-1}, Table 2) in comparison with experimental data (-5.2 ± 0.1 kcal mol^{-1}). The running average for both protocols is stable after ca. 20% of the simulation duration.

For OA-G6, (Figure 3.6C) the SOMD free energies rapidly converge to very similar values with the predictions or the other participants. The calculated binding free energies are consistent with YANK's predictions (-6.9 ± 0.1 kcal mol^{-1} vs -7.1 ± 0.1 kcal mol^{-1} respectively). These figures are in better agreement with experiment (-5.0 ± 0.1 kcal mol^{-1}) than the SAMPL6 SOMD free energies (-8.1 ± 0.2 kcal mol^{-1}).

Overall comparison of free energies estimated from the SAMPL6 and

SAMPLing protocols shows that averaging results over multiple starting host-guest structures improved agreement of predictions with experiment for OA-G3 and OA-G6 but not CB8. The differences in binding free energies computed by SOMD and YANK might be due to the differences in the simulation protocols.

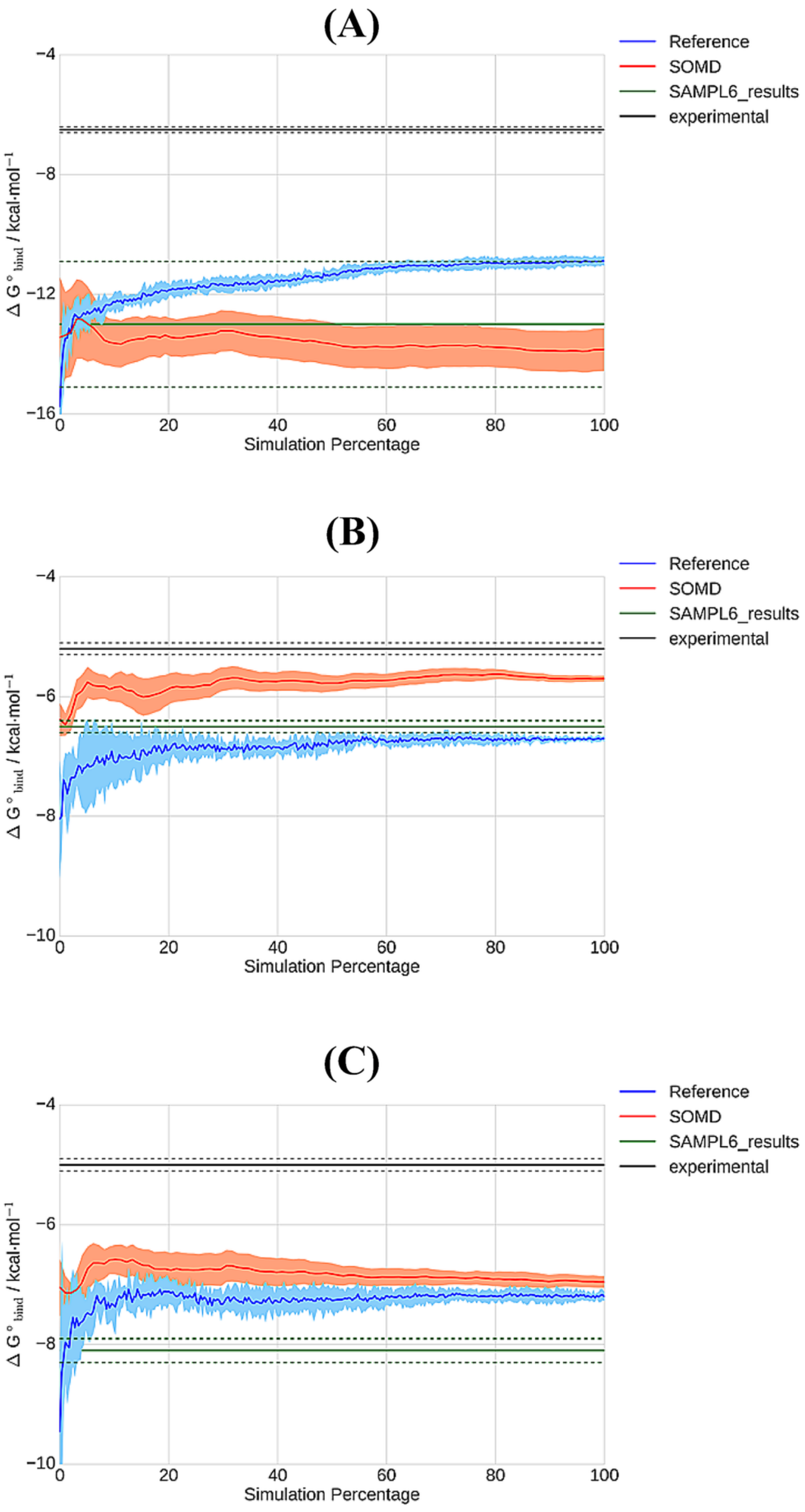


FIGURE 3.6: Comparison of standard binding free energies computed with SOMD (red) to SAMPLing reference values (blue) for CB8-G3 (A), OA-G3 (B) and OA-G6 (C). Bold lines denote the average free energy from the five different independent simulations started from different coordinates. Shaded areas denote 1σ. The experimental and SAMPL6 results are illustrated as black and green lines respectively, and the dotted lines denote 1σ.

3.9 Conclusions

Alchemical free energy calculations were applied to estimate standard binding free energies for 27 host-guests in the SAMPL6 competition. Protocols similar to that used in the SAMPL5 competition were implemented (*Model A/no-buffer* and *Model B/no-buffer*)[182], leading to results of comparable performance to SAMPL5 (SAMPL6 *Model B/no-buffer* R^2 ca. 0.6, MUE 3.9 kcal mol^{-1}, N=27 vs SAMPL5 *Model C* R^2 ca. 0.7 , MUE 3.4 kcal mol^{-1}, N=22). Additionally, an empirical correction term derived by a linear regression approach against SAMPL5 data was designed to correct for systematic errors in the free energy calculation protocol (*Model C/no-buffer*). This leads to significant improvements in mean-unsigned error but a slight decrease in correlation with the experimental data (MUE ca.1.4 kcal mol^{-1}, R^2 ca. 0.5). High accuracy predictions and correlations with experimental data were achieved for the two octa-acid hosts, but *CB8* proved more challenging, with significantly higher uncertainties in the computed binding free energies and poor correlation with the ITC measurements.

The influence of the modelled buffer on the computed free energies of binding was also investigated. The main finding is that explicit modelling of the buffer weakens binding of positively charged guests to *CB8* and enhances the binding of negatively charged guests to the two octa-acid systems. Overall the MUE for the dataset (*Model A* and *Model B*) drops by about 0.6 kcal mol^{-1} because the *CB8* binding energies are more in line with experimental data. However, this enhancement is also accompanied by a drop of ca. 0.2 in R^2. The empirical correction term derived against SAMPL5 data is incompatible with a protocol that models explicitly a buffer, presumably because no buffer was modelled in the SAMPL5 calculations.[183]

With respect to other SAMPL6 submissions, the results obtained with SOMD were promising and among the top performing models for the two octa-acid host-guest systems as judged by R^2 and MUE metrics. *CB8* proved challenging for most of the other softwares and methods used. SOMD *Model C/no-buffer* gave the lowest MUE values among all submissions (ca. 1.5 kcal mol^{-1}), but the predictive power was trivial (R^2 ca. 0.1).

Concerning the SAMPLing challenge, seven free energy methods were applied in three host-guest systems, namely OA-G3, OA-G6 and CB8-G3, that were parameterised with the same force-field. Their accuracy and their level of agreement were compared through an analysis framework that was devised by the organisers. Overall, this analysis showed that absolute binding free energy calculations can converge within reasonable computing time for these kinds of systems. However, this research highlighted significant and system-dependent discrepancies in the methods' convergence properties, that depend on both the free energy estimator and the sampling strategies used. In addition, this study illustrated that different methodological choices, software packages and/or details of the simulation, that should have trivial impact on the predictions, can introduce significant differences in the converged free energy estimates for the different methods ranging from 0.3 to 1 kcal mol^{-1}. Thus, there is a need for further investigation for the factors that contribute to some of these discrepancies.

Regarding our calculations, the OA-G3 and OA-G6 binding free energies computed from SOMD with the SAMPLing protocol were significantly different from those computed with SAMPL6 protocol (0.7 and 1.2 kcal mol^{-1} respectively). A standard practice in the Michel group is to estimate uncertainties in computed binding free energies from triplicate runs starting from the same input coordinates. This provides a reasonable estimate of the extent to which free energies are reproducible given a starting condition, but can also give an ambiguous impression of convergence. Where multiple reasonable poses can be produced, efforts are better spent evaluating free energies with simulations started from different input coordinates. Comparison of SOMD's free energies with the reference values (YANK) provided by the organizers yields a mixed picture, with a significant difference (CB8-G3, 3 0.7 kcal mol^{-1}), a moderate difference (OA-G3, 1 0.2 kcal mol^{-1}), and one trivial difference (OA-G6 0.2 0.2 kcal mol^{-1}). There are several differences between the two codes that could explain discrepancies, a notable one being a reaction-field treatment of long-range electrostatics (SOMD) versus PME (YANK). Other differences exist around the coupling of non-bonded and bonded interactions with the λ schedule,the treatment of soft-cores and electrostatic correction terms for charged guests. More systematic reproducibility studies on larger datasets need to be carried out to find the origins to the observed variability. Such efforts are important to validate the robustness and transferability of molecular simulation algorithms.

Chapter 4

Prediction of Absolute Binding Free Energies of Ligands for the Intrinsically Disordered Protein c-Myc

4.1 Introduction

4.1.1 Intrinsically Disordered Proteins

The field of structural biology has assumed for almost a century that the 3D structure of a protein was dictated by its amino acid sequence, and that a folded protein structure was necessary for biological function. However, in 1990s the discovery of proteins containing disordered regions linked to a biological function led many to question this dogma. The notion that proteins can be biologically active while remaining unstructured has become increasingly prevalent. In 1999, Dyson *et al.* introduced the term Intrinsically Disordered Proteins (IDPs) to describe a class of proteins that can be partially or completely unfolded, and yet biologically active.[230] A growing number of IDPs have been reported over the years, and are now classified in the database DisProt.[231] IDPs are composed of protein sequences that are unable to fold spontaneously into stable, well-defined globular three-dimensional structures but are dynamically disordered and fluctuate rapidly over an ensemble of conformations.[232–234] Proteins need typically to contain a disordered segment of ca. 30 or longer residues in its native state to be classified as IDPs. This disordered region is characterised by little or no tertiary and secondary structure under physiologic conditions. Disordered regions are typically depleted hydrophobic residues and contain more charged residues than structured regions.[235–237] An example of the 3D structures of a well-folded versus a disordered protein is illustrated in Figure 4.1.

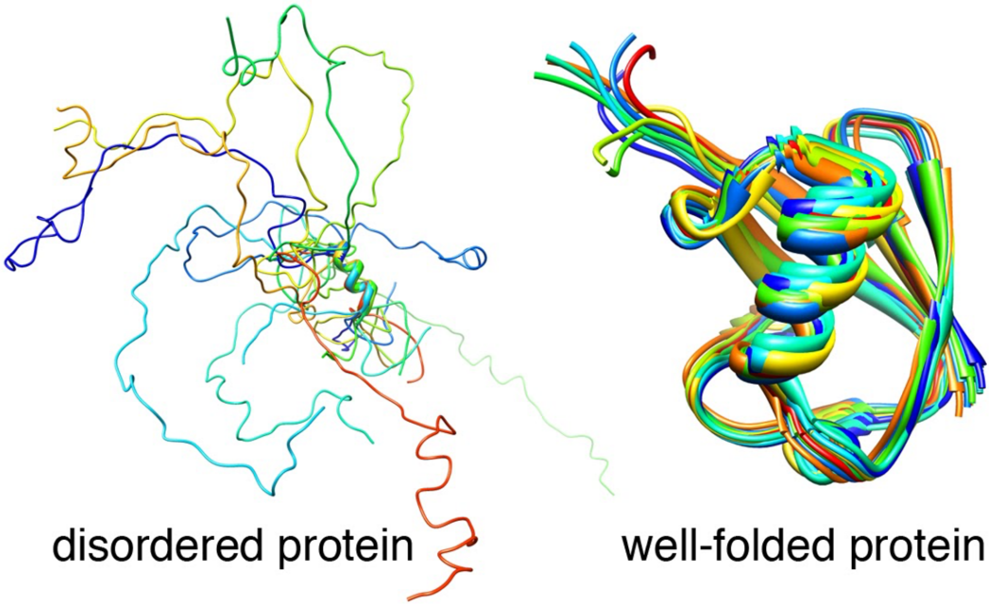


FIGURE 4.1: A schematic example of the 3D representation of a disordered versus a well-folded protein. Coloured tubes denote different conformational states the protein adopt in native conditions. Adapted from paper Cino *et al*.[238]

IDPs are highly abundant in nature. They are predicted to amount for 40% of eukaryotic, 25% of viral and 10% of bacterial proteins.[234] They participate in protein-protein interactions through a coupled-folding upon binding mechanism. This mechanism is characterized by high-specificity low-affinity complexes due to the high entropic cost of complex formation.[239] There are however some examples of IDPs that are not ordered upon complex creation.[240] In addition, IDPs can interact with multiple partners in one-to-many or many-to-one binding by changing shape to bind with different targets[241–243] and are though to exhibit intrinsic "conformational preference" for structures adopted upon binding to a protein partner.[244]

Thanks to their structural elasticity, they play vital roles in a plethora of cellular function such as signalling or transcription.[239, 245, 246] In addition, IDPs are involved in a variety of diseases such as cancer, cardiovascular disease, neurodegenerative disease, and diabetes.[247] This common implication of IDPs in the pathogenesis of various diseases led to the 'disorder in disorders' or D^2 concept,[248] according to which IDPs are richly involved in the development of numerous diseases due to their unique functional and structural properties. Therefore, such diseases may arise from the misregulation, misidentification of binding partners, missignaling and misfolding of the responsible IDPs.[235, 249–251] Thus, given their abundance and their biological importance there is a need for chemical agents that may control their function. However, until recently IDPs were considered as undruggable since their considerable flexibility is an inherent challenge for structure based drug design approaches. A major impediment is that this flexibility restricts the applicability of established structure-based methods such as NMR or X-ray crystallography to show in detail protein-ligand interactions. Additionally, little is known about the molecular driving forces that underpin IDP recognition, and how such principles can inform the design of man-made molecules that can effectively modulate the function of IDPs.

Yet some reports (Figure 4.2) have demonstrated inhibition of the biological functions of IDPs. These approaches take advantage of the heterogeneous nature of IDPs, which can consist of both ordered and disordered regions, and the ability of IDPs to interact with structured partners. One technique finds drug-like molecules that bind to the ordered domains of the IDPs and inhibit their biological functions. Another technique focuses on targeting critical regions of the IDPs, called molecular recognition elements (MoRE), that identify the structured binding partners and fold upon binding with them. After the structural determination of these complexes, conventional drug design methods can be employed to discover small molecules that mimic these regions and compete for the binding site with the ordered partner.[247],[252] In both instances, structure based drug design techniques are used to target the ordered proteins.

However, many protein-protein interactions consist of two IDPs whose complex cannot be solved in isolation. Even in cases where the structure of the complex is ordered, there may be not apparent pockets that drug molecules can readily bind. Thus, a third approach involves the direct targeting of the functional disordered state of the IDPs. Several IDPs can functionally misfold through non-native intramolecular interactions of their sticky preformed binding elements that form a non-interacting or a less-interacting cage. This mechanism takes advantage of the structural elasticity of these proteins and their ability to morph into differently shaped bound configurations to prevent them from binding to non-native partners.[240] Therefore, this concept can be exploited to find small molecules that stabilise functionally inactive conformational ensembles. The idea is to discover drug molecules that could stabilise the disordered state of an IDP in an structure different from the structure adopted in the complex with its binding partner, thus inhibiting biological important interactions that require coupled folding and binding. This method has the advantage that it does not need high resolution 3D data for the binding partner but on the other hand it has the drawback that existing structure based drug design methods cannot be applied in this case.[253]

In addition, small molecules can be used to target aggregation structures or to promote the formation and stabilization of non-amyloidogenic and non-toxic oligomeric or monomeric species.[254–256] A recent example of this strategy is a series of drug molecules called ‘molecular tweezers’ that disturb the oligomerization and aggregation processes of several proteins such as amyloid-protein (Ab) and a-synuclein.[257],[258] Such molecules specifically bind to lysine residues by encapsulating them into electron-rich torus shaped cavities decorated with two rotatable peripheral anionic phosphonate groups.[259] These tweezers were shown to inhibit the toxicity of these proteins and could be used to treat other protein misfolding diseases, such as Alzheimer’s and Parkinson’s disease.

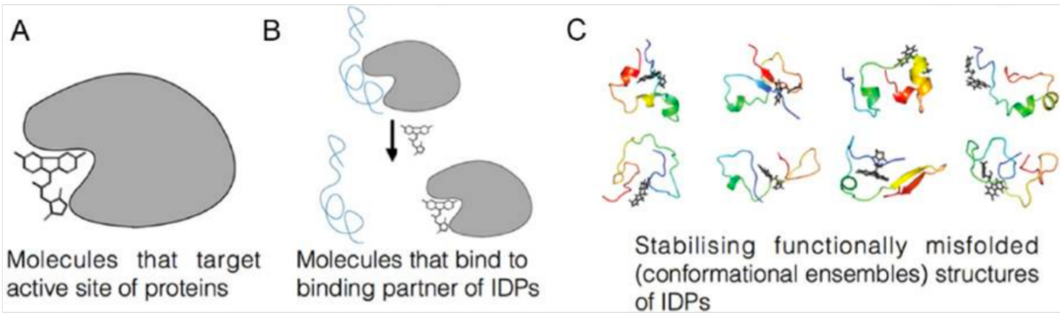

A
B
C
Molecules that target active site of proteins
Molecules that bind to binding partner of IDPs
Stabilising functionally misfolded (conformational ensembles) structures of IDPs

FIGURE 4.2: Different approaches to disorder-based drug design. (A) Structure-based drug design method that targets active sites of ordered proteins using drug molecules. (B) Inhibition of protein-protein interactions between an IDP (blue) and its binding partner (grey) by a small molecule (black) that can bind to the ordered structure of the binding partner. (C) stabilisation of inactive disordered states by a small-molecule which inhibits the biological functions of the IDP. Adapted from Jin *et al* and Kumar *et al*.[260],[261]

4.1.2 The Oncoprotein c-Myc

One striking example of the direct targeting approach is provided by the the oncoprotein c-Myc. This 65 kDa nuclear phosphoro-protein consists of 439 amino acids and contains an an N-terminal transactivation domain (TAD) and an 88-amino-acid C-terminal bHLHZip (basic helix-loop-helix leucine zipper) domain.[262] This IDP belongs to the Myc family of transcription factors and serves as a key regulator of numerous genes that are involved in diverse cellular processes such as cell proliferation, differentiation, metabolism, adhesion, apoptosis, maintenance of cell size, genomic integrity, and angiogenesis. c-Myc is overexpressed in most human cancers including breast cancer, colon cancer, cervical cancer, small-cell lung carcinomas, osteosarcomas, glioblastomas, melanoma, and myeloid leukaemias and is often considered the 'holy grail' in cancer therapy.[247, 263–265] Indeed several studies have shown that successfully targeting c-Myc can lead to cell-cycle arrest, apoptosis, re-differentiation of tumor cells, tumor vascular degeneration, and finally tumor regression.[265–272]

In order for c-Myc to act as a transcription factor and thus play its important biological role it must heterodimerize with the basic-Helix-Loop-Helix-Leucine zipper (bHLHZip) domain of partner protein Max which lacks a transactivation segment. The c-Myc/Max heterodimers can be thought of as microscopic scissors that recognise the DNA double helix (Figure 4.3). In this arrangement they are able to recognise DNA response elements such as E-boxes (enhancer boxes).[273–276] Therefore it has been shown that the disruption of this interaction is a possible anticancer strategy.[245]

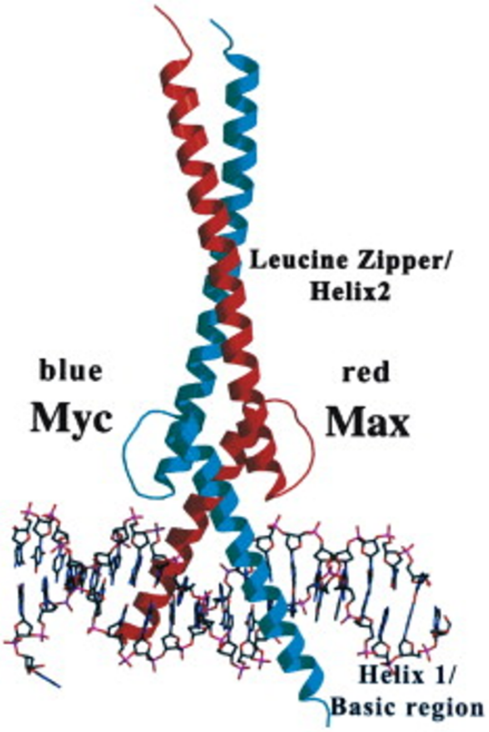


FIGURE 4.3: Structure of the bHLHZip domains of Myc (cyan) and Max (red) in complex with a DNA sequence. Adapted from Turner *et al* [276].

An obvious approach would be to specifically inhibit the formation of the heterodimer, since a crystal structure of the c-Myc/Max together with the DNA is available. However, the application of structure-based drug design approaches is challenging as both c-Myc and Max are highly disordered in their unbound states and they fold into a helical coiled coil only when they interact with DNA.[240, 247, 252] Ultimately, there is a lack of potential binding sites in this heterodimer.[277] This is the reason why most of the inhibitors obtained so far were found by high-throughput screening rather structure-based approaches.[278]

The first successful HTS attempt was performed by the Vogt lab at the Scripps Research Institute. They screened 7000 compounds from a combinatorial library and identified two molecules, IIA4B20 and IIA6B17. These peptidomimetics suppressed c-Myc dependent cell growth. Subsequently, Vogt and Boger *et al.* replaced the isoindoline core of the these compounds with a racemic, trans-3,4 dicarboxylic acid core which led to the discovery of mycmycin-1 and mycmycin-2.[279] These second generation c-Myc inhibitors showed a 10-fold stronger inhibition than IIA6B17. In addition, they demonstrated greater selectivity that the first generation of Myc inhibitors as they did not inhibit the oncogenic transcription factor c-Jun.[280] High-throughput screening of additional libraries paved the way for the discovery of new small molecule inhibitors of the c-Myc/Max heterodimerization. A fluorescence resonance energy transfer (FRET) assay was performed by Xu *et al.* on a library of hydrophobic and planar compounds (a so called "credit card" library) taking advantage of the observation that most protein-protein interactions have hydrophobic interfaces. Two compounds, NY2267 and NY2280 showed disruption of the c-Myc-Max dimerization, inhibition of specific DNA binding, and inhibition of oncogenic transformation. However, they did not provide the desired selectivity as they also showed similar inhibition of c-Jun.[281] The screening of combinatorial libraries by Kiessling *et al.* led to the discovery of a promising pyrazolo[1,5-a]pyrimidine scaffold for c-Myc inhibition. Mycro3, that was built upon the two parent compounds Mycro1 and Mycro2, was the first inhibitor that also disrupted the c-Myc-Max DNA binding. Furthermore, Mycro3 demonstrated a strong and selective inhibition of c-Myc/Max heterodimerization *in vitro* and improved pharmacokinetics and bioavailability.[282] Finally, three more recent inhibitors where discovered using the same strategy. KJ-Pyr-9 exhibited a direct intracellular binding to c-Myc with nanomolar affinity (K_D of 6.5 nM),[283] sAJM589 bound potently to the Myc bHLHLZ domain with an IC_{50} of 1.8 μM,[284] and finally MYCMI-6 (NSC354961) demonstrated a potent inhibition of c-Myc/Max dimerization with a K_D of 1.6 μM, based on surface plasmon resonance (SPR) assays.[285],[286]

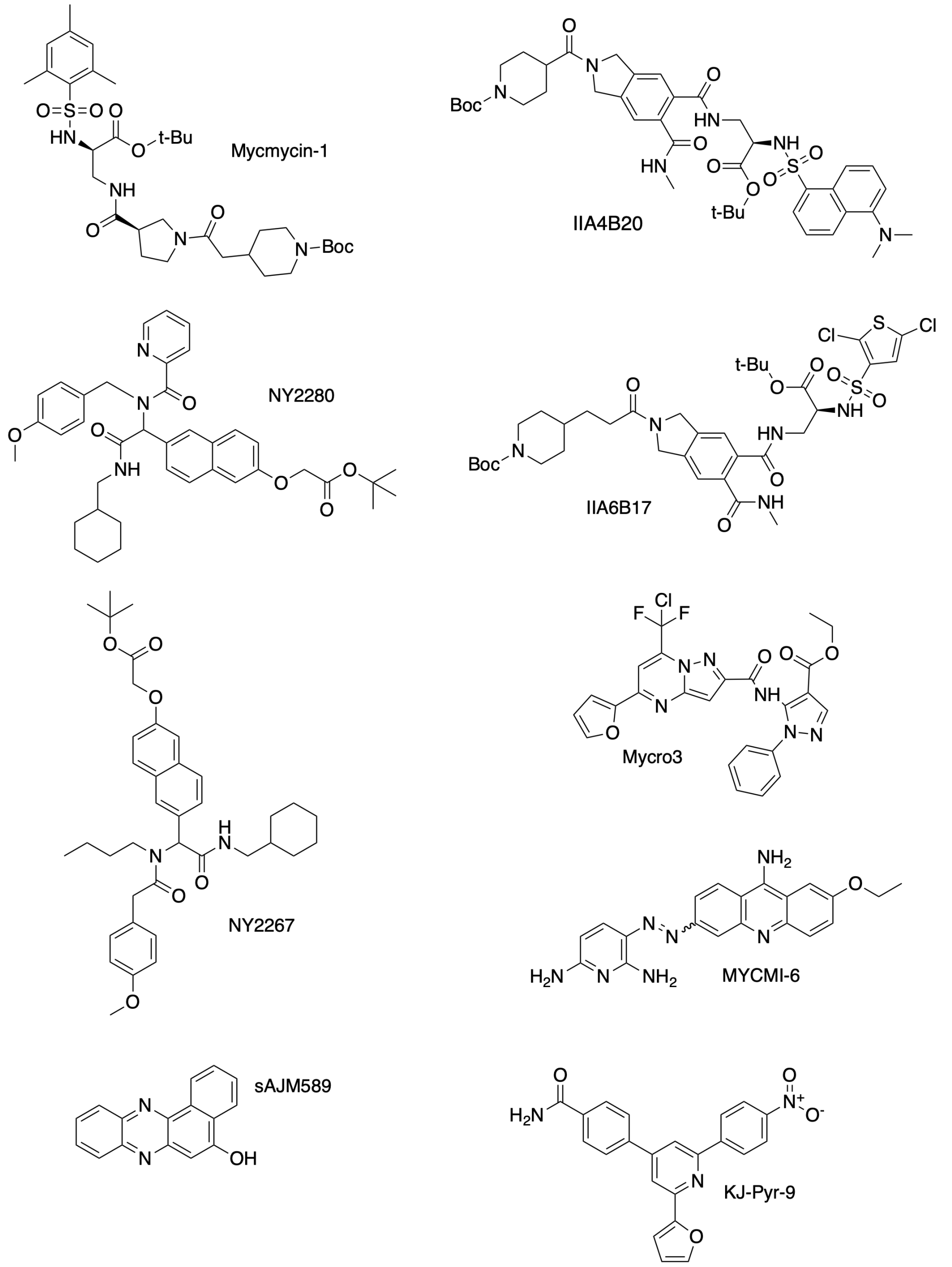
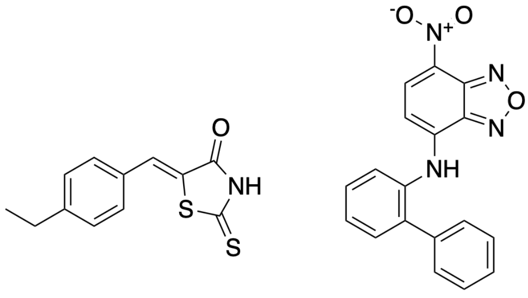


FIGURE 4.4: Chemical structures of selected small molecule inhibitors that target c-Myc/Max dimerisation. [247],[286]

Among all of the known inhibitors only 10058-F4 and 10074-G5 (together with KJ-Pyr-9) demonstrated complete selectivity towards c-Myc/Max dimerization and shown to bind directly to c-Myc by Circular Dichroism (CD), NMR experiments.[262],[252] In 2003, Prochownik's group screened 10,000 compounds from the Chembridge DIVERSE combinatorial library and found three molecules were able to selectively inhibit the c-Myc transcriptional function and decrease the fibroblast growth.[269] The rhodanine-based compound 10058-F4 inhibited the proliferation of HL60 cells with an IC$_{50}$ value of 41.1 μM, while 10074-G5 inhibited the growth of these human promyelocytic leukemia cells that overexpress c-Myc with an IC$_{50}$ value of 22.5 μM.[287] Subsequent fluorescent polarization assays indicated that 10058-F4 binds to c-Myc with a K$_D$ of 5.3 μM and 10074-G5 binds to the oncoprotein with a K$_D$ of 2.8 μM. Additional surface plasma resonance (SPR) experiments by Müller *et al* determined the direct binding of 10058-F4 and 10074-G5 with K$_D$ values of 39.7±8.1 μM and 31.7±24.9 μM, respectively.[288] Finally, Heller *et al* examined the potency of 10058-F4 with Isothermal Titration Calorimetry (ITC) and with van't Hoff analysis using fluorescence titration experiments at different temperatures. They did not observe any binding with ITC because of low heat of binding, but they measured a binding free energy of -27.6 ± 8.5 kJ/mol at 25 C° binding with a van't Hoff analysis. Thus, they concluded that entropic contributions are a key factor for the binding of 10058-F4 to the oncoprotein c-Myc.[289] The chemical structures of the two promising inhibitors are shown in Figure 4.5.



10058-F4 **10074-G5**

FIGURE 4.5: Illustration of the chemical structures of 10058-F4 and 10074-G5.

Despite the rising number of c-Myc/Max inhibitors, it is also important to elucidate the potential binding sites of the intrinsic oncoprotein. c-Myc$_{402-412}$ and c-Myc$_{363-381}$ were identified as the binding sites of 10058-F4 and 10074-G5 respectively, through point mutations and deletions of short, synthetic peptides of c-Myc, followed by CD and NMR assays.[262] In addition, Hammoudeh *et al* proved that the binding of these compounds to these two sites can happen simultaneously and independently. Finally, a model of 10058-F4 bound to c-Myc$_{402-412}$ was proposed on the basis of NMR experiments. This region is located at the interface between the H2 and Zip region in the c-Myc-Max dimer and forms a hydrophobic cluster of Tyr402, Ile403, Leu404, Val406, Ala408 in the c-Myc$_{402-412}$10058-F4 complex.[252] Therefore, using these findings as a starting point, further studies can be performed to better understand the binding mechanism of these molecules with c-Myc

4.1.3 Previous work from the Michel group and scope for further improvements

Intrigued by this opportunity, Michel and Cuchillo employed molecular dynamics and bias-exchange metadynamics simulations to provide new insights about the mechanisms of molecular recognition between the small molecule 10058-F4 and c-Myc. They reported that the ligand does not have a dominant binding mode but interacts with multiple binding sites through weak and non-specific interactions. Moreover, the compound made preferential contacts with the most hydrophobic region of this sequence, a result that was in agreement with Hammoudeh *et al*.[290],[252] Therefore, this study has highlighted the lack of specificity between 10058-F4 and its target as well as the difficulty to locate possible binding sites for c-Myc.

A first successful example of a general computational approach for targeting c-Myc target has been reported by Yu *et al*. They performed extensive MD sampling of the c-Myc$_{370\text{˘}409}$ region to extract an ensemble of conformations that were used for binding site identification and multi-conformational

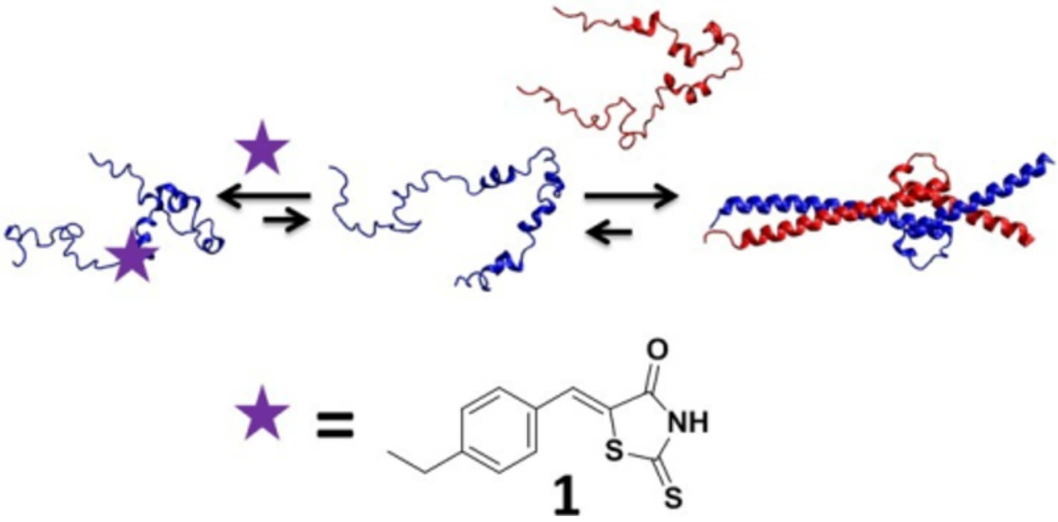


FIGURE 4.6: The small molecule 10058-F4 (purple star, 1) disturbs the heterodimerization of c-Myc (blue), and Max (red), by preserving conformations in c-Myc that are unsuited with c-Myc/Max dimerization. Adapted figure from Cuchillo *et al.*[290]

molecular docking.[260],[121] However, it remains important to further establish the mechanisms by which small molecules can interact with c-Myc. This is a really challenging task through classical experimental approaches such as X-ray crystallography and NMR, because of the highly disordered nature of c-Myc. In contrast, MD simulations can be performed to study and understand the molecular recognition mechanism between a ligand and c-Myc, since these interactions are often linked with rapid conformational changes.[291]

The aim of this project is to characterise the interaction between 10058-F4 and c-Myc using two MD-based strategies. The first approach assesses the the ability of the absolute binding FEP protocol employed in Chapter 3 to reproduce the binding affinity and binding site preference of two well-characterized c-Myc binders, 10058-F4 and 10074-G5, on the 402-412 Myc fragment. It was shown in Hammoudeh *et al* that 10058-F4 prefers to bind in this segment, while 10074-G5 binds to a different region (363-381) of c-Myc.[252] The second approach uses long unbiased MD simulations to reversibly simulate binding/unbinding of 10058-F4 to c-Myc$_{402-412}$ and analyze the resulted trajectories using MSMs.[292],[59] Overall we aim to test for the consistency of the binding free energies computed by these two approaches, and to gain insights in the binding mechanism of 10058-F4 to c-Myc.

4.2 Computational workflow

4.2.1 MD simulations of c-Myc-ligand complexes

For the MD simulations, the binding pose from Hammoudeh *et al.* was chosen as a starting point for the c-Myc$_{402-412}$/10058-F4 complex. In addition, for the c-Myc$_{402-412}$/10074-G5 complex, the ligand was docked to the same Hammoudeh *et al.* pose as 10058-F4 using Cresset's molecular modelling package Flare.[162] The docking grid was centered on 10058-F4 bound to c-Myc and the whole peptide was used as a receptor. The lowest energetic pose from this complex was chosen as a starting point for the MD simulations.

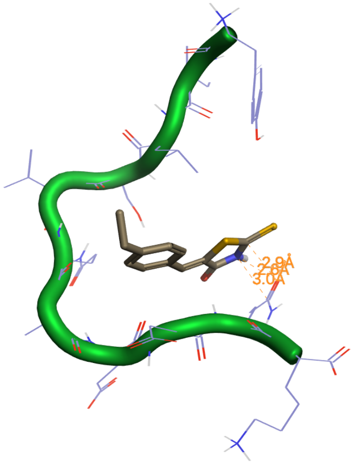


FIGURE 4.7: Depiction of the starting point of c-Myc$_{402-412}$/10058-F4 complex for the MD simulations. The thiazolidinone ring forms hydrogen bonding interactions with the side chain of Gln$_{411}$. [252]

Two sets of equilibrium MD simulations were performed for this study. The first set was implemented to derive the most stable pose as an initial conformation for the absoluteFEP protocol for the two protein-ligand complexes. For this purpose, the input files used for these simulations were created using FESetup1.2.1 software.[167] Proteins were parameterised using ff14SB Amber force-field[169], while GAFF2 parameters[170],[171] that use AM1-BCC charges[172] were assigned to the ligands. Both systems were solvated in a rectangular box with TIP3P waters[173] with a minimum distance between the solute and the box of 12 Å. Counter ions were also added to neutralise the total net charge.

For the equilibration protocol, energy minimization of the entire system was implemented with 1000 steps of steepest gradients, using sander. Then, an NVT protocol for 200 ps was performed at 298K, followed by an NPT equilibration for further 200 ps at 1 atm. Eventually, a 2 ns MD simulation in an NPT ensemble was run with sander to reach a final density of 1 g cm^{-3}. Simulations of each protein-ligand complex were run for a duration of 500 ns using SOMD software in an NPT ensemble. Temperature control was maintained by an Andersen thermostat with a coupling constant of 10 ps^{-1}. Pressure control was achieved with a Monte Carlo barostat that attempted isotropic box edge scaling every 25 fs. A 10 Å atom-based cutoff distance for the non-bonded interactions was used, using a Barker Watts reaction field, with dielectric constant of 78.3. The final coordinate files were retrieved with cpptraj.

The second set of MD simulations consist in three different long MD simulations to build a Markov state model of the c-Myc$_{402-412}$/10058-F4. Each simulation was performed with a different force field.

The first set of input files for the protein-ligand complex were generated with the same method and force-fields as for the first set of MD simulations. The same equilibration protocol was employed and the final coordinate file was obtained with cpptraj. A 20 μs long MD simulations for the protein-ligand complex was run using the SOMD software (revision 2019.1.0) in the NPT ensemble at 300 K and 1 atm. A 2 fs timestep was used and all the bonds involving hydrogens were constrained. Temperature control was maintained by an Andersen thermostat with a coupling constant of 10 ps^{-1}. Pressure control was achieved using a Monte Carlo barostat. Periodic boundary conditions were used with a 10 Å atom-based cutoff distance for the non-bonded interactions together with a Barker Watts reaction field with dielectric constant of 78.3 for the electrostatic interactions.

For the second simulation the ff14IDPSFF Amber force-field[293] was selected for c-Myc$_{402-412}$ as it is a specific force-field for IDPs, GAFF2 parameters[170],[171] with AM1-BCC partial charges for the ligand through the LEaP module in the Amber 17 suite.[168] The model was then solvated in a rectangular box of TIP3P water molecules and charge neutrality was enforced through the addition of the necessary counter ions. The input coordinates were energy minimised using 5000 steps of steepest gradients with heavy protein atoms were position-restrained with a force constant of 1000 kJ mol$_1$ nm$_2$. The system was then equilibrated for 100 ps using an NVT ensemble and the same restraints as in the previous step. Finally 100 ps of NPT ensemble at 1 atm were performed to reach a final density of about 1 g cm^{-3}. Next the software GROMACS 5.0.5[225] was used to perform a 20 μs long MD simulation for the protein-ligand complex in the NPT ensemble at 300 K and 1 atm. A 2 fs timestep was used and LINCS[294] algorithm was employed to constrain bonds involving hydrogen. Temperature control was maintained at 300K with a stochastic Berendsen thermostat.[295] and pressure was achieved using a Parrinello-Rahman barostat.[296] Electrostatic interactions were handled using Particle Mesh Ewald with a real space cutoff of 10 AÅ and a Fourier grid spacing of 1.6 Å. Van der Waals interactions were handled using Lennard-Jones with a cut-off of 10 AÅ.[297]

The third force-field tested was Charmm36m which has been parameterised for IDPs,[298] GROMACS 5.0.5 package[225] was used to prepare the third set of the input files. The general Charmm force field[299] was selected for the ligand. The model was then solubilised in a rectangular box with TIP3P waters[173] with a box length of 12 AÅ away from the edge of the solute. In addition, counter ions were added to neutralise the total net charge. A similar equilibration and production protocol as for the previous setup was followed to produce a 20 μs MD simulations with the software

GROMACS 5.0.5 [225].

4.2.2 Double decoupling protocol

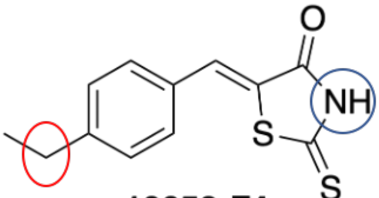
Alchemical free energy simulations were performed using a double decoupling protocol implemented in the SOMD software. Details of the protocol were given in Chapter 3. For the two protein-ligand complexes, both complex and solvated phase *discharging* step were run with nine equidistant λ windows and 16 λ windows (0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.400, 0.45, 0.50, 0.55, 0.60, 0.70, 0.85, 1.00) were employed for the *vanishing* step, both in bound and free phase.

Each lambda value was simulated for a duration of 10 ns with SOMD in the NPT ensemble. Temperature control was achieved with an Andersen Thermostat with a coupling constant of 10 ps^{-1}.[48] Pressure control was maintained by a Monte Carlo barostat that attempted isotropic box edge scaling every 100 fs. A 12 Å atom-based cutoff distance for the non-bonded interactions was used, using a Barker Watts reaction field with dielectric constant of 78.3.[223] In the bound phase the restraints parameters of eq. 3.1 were: R_{ji} = 7 Å, D_{ji} = 2 Å and k_{ji}= 10kcal mol^{-1} Å^{-2}. The alpha carbons (C_a) of residues Leu_{404} and Gln_{410} were chosen as the restraint set of the host atom, while the central carbon atom of 10058-F4 and the central nitrogen atom of 10074-G5 were the corresponding guest atoms.

Free energy changes were estimated with the multistate Bennet acceptance ratio method as implemented in the Sire app *analysefreenrg*.[50] To achieve a more robust estimation of free energies, each simulation was repeated three times, using different initial velocities drawn from the Maxwell-Boltzmann distribution and statistical uncertainties are reported as one standard error of the mean.

4.2.3 Markov State Model's protocol

The resulting pool of trajectories from the second set of MD simulations was used to construct a MSM for the three different force-fields using the pyEMMA 2.3.0 software package.[300] Three different features were used to cluster the MD simulations to build MSM models. All features involved distances between the circled atoms of 10058-F4 (Figure 4.8) and the C_a atoms of the c-Myc peptide.



10058-F4

FIGURE 4.8: Structure of 10058-F4. The two atoms of the molecule that are circled were chosen to measure distances between the ligand and the alpha carbons of the c-Myc peptide. These distances were used as molecular features for the MSM models.

The first metric required the calculation of eleven distances between the nitrogen atom of the ligand and the CAs of each amino acid of the c-Myc peptide in each snapshot (Metric 1). Then, dimensionality reduction was performed using TICA to construct a lower dimensionality representation that could captured the variance of this 11-dimensional space.[55] However, TICA retained ten out of eleven dimensions to explain 95 % of the total variance of the system. Thus, we decided to use the original eleven dimensions for each snapshot, as there was not significant dimensionality reduction using TICA.

The second metric used only the shortest distance between the nitrogen of the ligand and the CAs of each amino acid of the oncoprotein for every snapshot (Metric 2). The third metric used the shortest distance between either the nitrogen or the carbon atoms highlighted in figure 5.8 and the alpha carbons of each residue at each snapshot (Metric 3). For the last two metrics we did not perform reduction of the dimensional space as the initial feature space consisted of only one dimension.

Subsequently, k-means clustering using 75 clusters was performed to discretize the trajectories and obtain microstates for the MSM construction. Implied timescales (ITS) of the dominant eigenvectors were calculated for each metric. All of the metrics have similar behavior compared to their slowest processes. From these plots, we concluded that a lagtime of 200 ps together with the default parameters of pyEMMA should be chosen to estimate the MSM transition matrices using the Bayesian MSM option of pyEMMa. The validity of these models were tested with the Chapman-Kolmogorov test (CK test) where the full transition probability matrix T was coarse-grained into 2 metastable states.[59]

Having ensured that the dynamics in the space of the 2 metastable states was Markovian, we performed spectral clustering using PCCA++ algorithm to coarse-grain the microstates into 2 metastable states.[64] The two states were identified as bound and unbound based on the maximum distance between the ligand and the protein in this state. In addition, the stationary probabilities (π) of the two metastable states were calculated by summing over the populations of the 75 microstates. The Mean First Passage Times (MFPT) between the two states were estimated from the Bayesian MSM using pyEMMA.

The standard binding free energy of the ligand was estimated using equation 4.1.

$$\Delta G^{\circ}_{\text{bind}} = -k_{\text{B}}T \ln\left(\frac{\pi^{bound}}{\pi^{unbound}}\right) - k_{\text{B}}T \ln\left(\frac{V^{\text{unbound}}}{V^{\circ}}\right), \tag{4.1}$$

where k_{B} is the Boltzmann constant, T is the temperature in Kelvin, π accounts for the stationary probabilities of the bound and unbound macrostates. The second term corrects for the volume of the unbound state in the simulation box being different from the standard volume conditions for a 1M dilute solute (V° = 1660 Å^3/mol).

To determine V_{bound}, the volume of space available to the ligand in the unbound state, we computed the average distance between the center of masses of the ligand and the protein from 1000 snapshots sampled from the bound macrostate. We then estimate the bound volume V_{bound} as the volume of a sphere with radius equal to this average distance. We also used cpptraj to compute the average volume V_{total} of the simulation box from these 1000 snapshots. The volume of the unbound state was then taken as the difference between V_{total} and V_{bound}.

Finally, we also computed the rate constants k_{on} and k_{off} for the binding process by using the calculated MFPTs between bound and unbound states. Assuming first order reactions, the relation between the rates and the corresponding MFPTs are provided by the following equations[301]:

$$k_{\mathrm{off}}^{-1} = MFPT_{\mathrm{off}}\,, \tag{4.2}$$

$$k_{\mathrm{on}}^{-1} = MFPT_{\mathrm{on}}C_{\mathrm{compound}}\,, \tag{4.3}$$

where C_{compound} is concentration of the ligand in the simulation box.

4.3 Results

4.3.1 Binding free energies from the double decoupling protocol

The established FEP protocol from Chapter 3 was employed to reproduce the binding affinity and binding site preference of two known c-Myc binders to the 402-412 Myc fragment. The binding free energies computed from the absoluteFEP protocol are provided in Table 4.1. Overall, standard state corrections are similar between the two ligands, but we can observe large differences between "discharge" and "vanish" steps of the protocol in free and bound phase. The protocol yielded quite reproducible results between the three independent runs, but the results do not show 10058-F4's preference for binding at this region. The FEP calculations predict that 10074-G5 binds more favorably to c-Myc$_{402\text{-}412}$ (-4.7±0.7 kcal/mol) than 10058-F4 (-0.8±0.2 kcal/mol) which contrasts with the previous reported experimental results.

A possible reason for this inconsistency could be the force-field parameters of 10058-F4. In 2017, Heller *et al* reported a custom parameterization of the ligand as they observed proved that GAFF was poorly representing the torsional energetics of this ligand. The GAFF and Heller's force field terms of the dihedral angle of 10058-F4 are shown in Figure 4.9.[289]

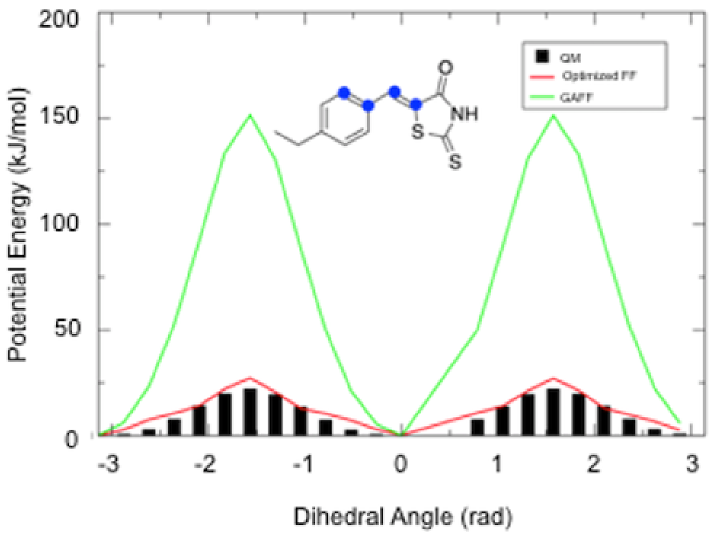


FIGURE 4.9: Parameterization of the force field term of the dihedral angle of 10058-F4 that is highlighted with four points in blue circles. The green and red lines represent the potential energies computed by GAFF and Heller's parameterization of the force field respectively. Black bars show the potential energy function calculated using quantum mechanical calculations at RB3LYP/6-311+G(d,p) level of theory. Adapted from Heller *et al.*[289]

Thus, we repeat the absoluteFEP protocol using the customized parameters for 10058-F4 that had been kindly provided by Dr. Heller. The most significant change was the 6-fold difference in $\Delta G_{bounddischarge}$ (71.67 ± 0.12 for GAFF versus 18.99 ± 0.08) and in $\Delta G_{freedischarge}$ (72.09 ± 0.01 for Amber force field versus 19.25 ± 0.01 for custom parameterisation) between the two simulations. Unfortunately, the binding free energy of the molecule was only slightly more negative (-1.3± 0.2 kcal/mol).

One possible reason for the poor computed energetics could be that the conformation of the c-Myc peptide was not representative of the dominant binding mode observed experimentally. To test for this we carried out 500-ns long MD simulations of the c-Myc/10058-F4 and c-Myc/10074-G5 complexes. Subsequently, we performed clustering with k-means algorithm and the RMSD of the ligand versus the protein as a metric using *cpptraj*. The most representative binding pose for each complex (Figure 4.10) was then selected and used as the initial conformation for the absoluteFEP protocol.

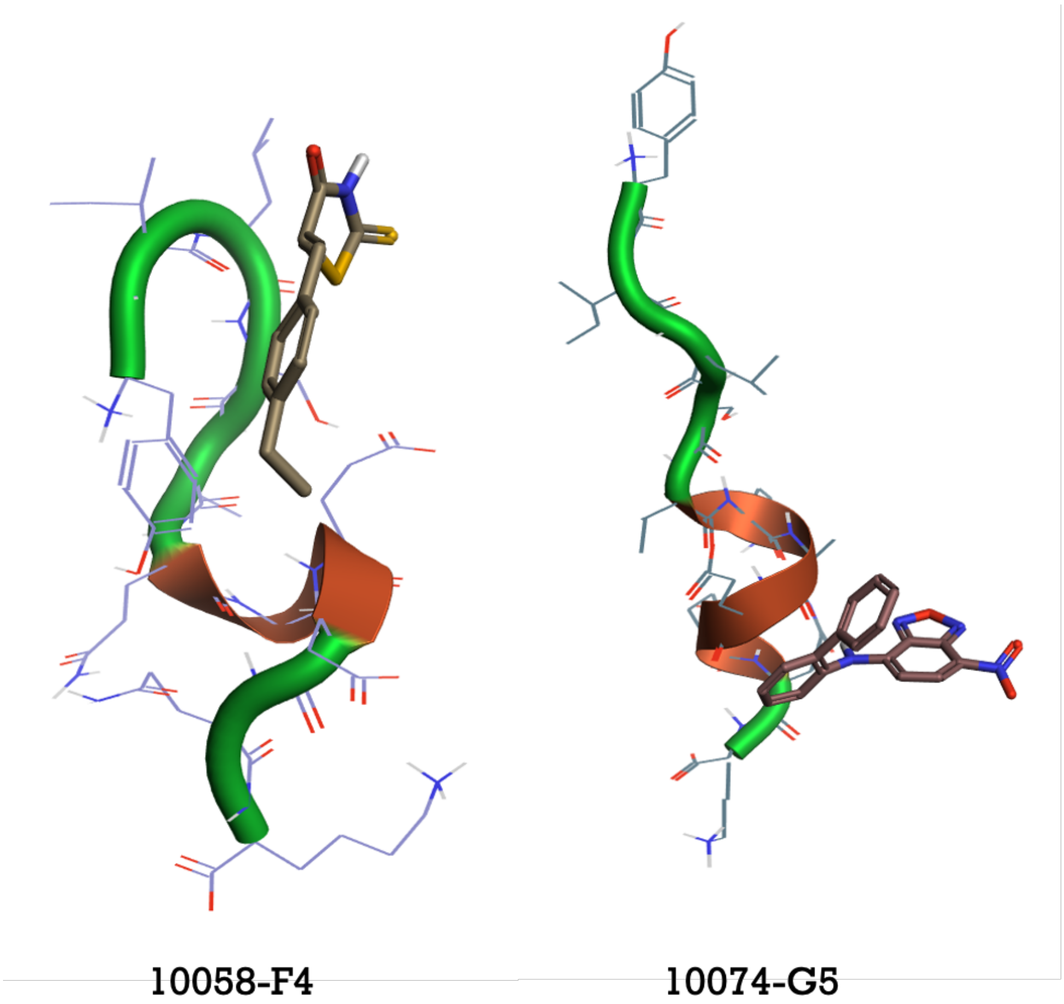


FIGURE 4.10: Depiction of the new, more stable binding poses of the two protein-ligand complexes used for the new runs of the absoluteFEP protocol.

The binding free energies computed using those structures were more positive than seen previously for both 10058-F4 (0.9±0.4 kcal/mol) and 10074-G5 (-1.4±0.3 kcal/mol). Furthermore, there was limited evidence of a dominant binding mode in the MD simulation, with the compounds reversibly binding/unbinding several times. This suggested challenges for the FEP protocol that affords limited sampling per window. We therefore considered next the evaluation of binding free energies using longer MD simulations.

AbsoluteFEP protocol for 10058-F4						
c-Myc	$\Delta G_{bounddischarge}$	$\Delta G_{boundvanish}$	$\Delta G_{freedischarge}$	$\Delta G_{freevanish}$	$\Delta G_{restraint}$	$\Delta G^{\circ}_{ModelC}$
Hammoudeh pose	71.67 ± 0.12	-4.06 ± 0.08	72.09 ± 0.00	-6.76 ± 0.01	-1.51 ± 0.26	-0.77 ± 0.21
Heller's parameters	18.99 ± 0.08	-4.21 ± 0.16	19.25 ± 0.00	-6.87 ± 0.07	-1.08 ± 0.12	-1.32 ± 0.18
Clustering pose	76.57 ± 0.07	-5.36 ± 0.36	77.47 ± 0.01	-6.56 ± 0.02	-1.21 ± 0.05	0.91 ± 0.39
AbsoluteFEP protocol 10074-G5						
Hammoudeh pose	52.55 ± 0.21	-16.90 ± 0.47	53.15 ± 0.02	-23.59 ± 0.04	-1.37 ± 0.04	-4.73 ± 0.68
Clustering pose	52.08 ± 0.07	-18.09 ± 0.31	53.28 ± 0.02	-21.47 ± 0.10	-1.51 ± 0.26	-1.38 ± 0.26

TABLE 4.1: Results from all the different absoluteFEP protocol simulations for 10058-F4 and 10074-G5. Energies are reported in kcal/mol

4.3.2 MSM results

Extensive 20 microsecond long MD simulations of the c-Myc$_{402-412}$/10058-F4 complex were conducted using three different force-field parameter sets. The rationale behind these simulations was to examine the behaviour of the c-Myc-ligand complex when protein force-fields developed specifically for IDPs are used. Three different metrics that are dependent on the distance between the protein and the ligand were used to construct three MSM models for each force-field from the simulation data using pyEMMA software. The ten slowest ITS for each metric and each force-field were plotted for a range of lag times, τ. The corresponding plot for the second metric is illustrated in Figure 4.11, while the other two plots are provided in the Appendix.

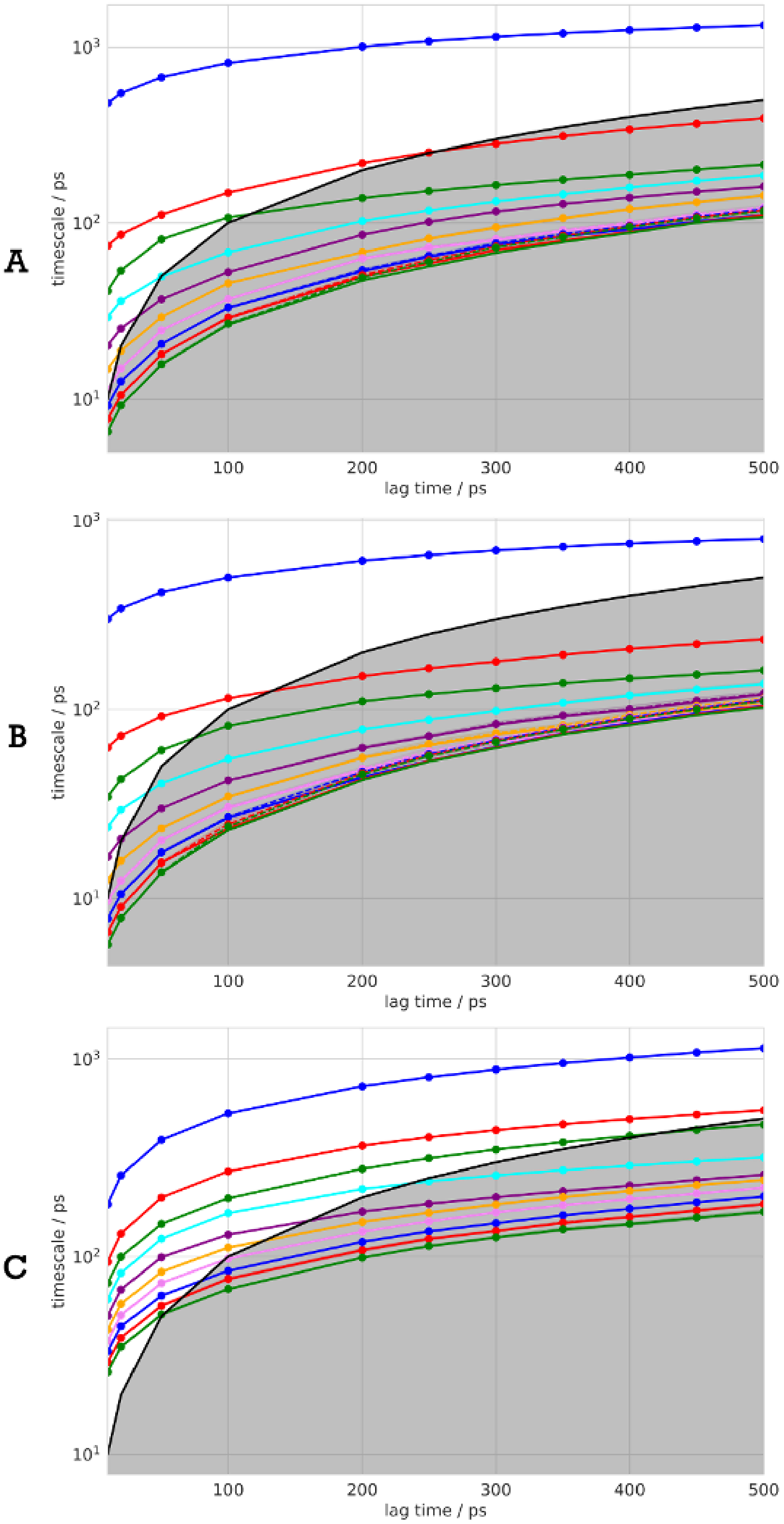


FIGURE 4.11: Implied time scales plots of the second metric (shortest distance between the nitrogen of the ligand and the CAs of each amino acid of the oncoprotein) for the three force fields, **A** AmberIDP, **B** Charmm36m, **C** FF14SB. Different colors indicate the slowest processes of the system during the MD simulations.

The lag time chosen for all analysis was 200 ps. The timescales haven't fully leveled off, and the model is only approximately Markovian. The validation of the Markovianity of the different models was performed through CK tests. The resulting plots for the second metric is depicted in Figure 4.12, while the other plots are given in the Appendix.

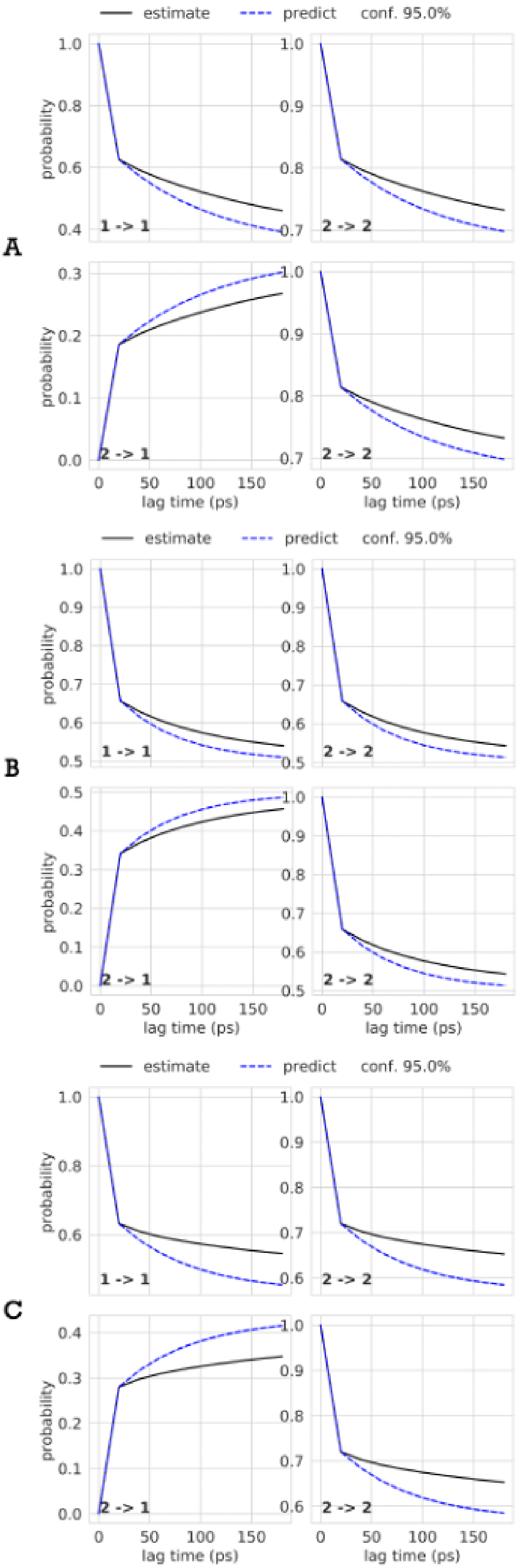


FIGURE 4.12: Chapman-Kolmogorov test plots of the second metric used for the three force fields, **A** AmberIDP, **B** Charmm36m, **C** FF14SB.

The CK tests for the second metric show a small deviation from the kinetic behavior of the system on longer timescales. Thus, a lag time of 200 ps is an appropriate choice to predict the long-timescale behavior of the three systems. After successful statistical validation of the models, a two macrostate MSM model was constructed for each simulation using PCCA++. The grouping turned out to distinguish different protein-ligand conformational states and separate them as bound and unbound based on the distance between the ligand central atom, and the CAs of c-Myc. The probability distribution of distances for the second metric in bound and unbound states is depicted in the Figure 4.13, while similar plots for the other metrics are provided in the Appendix. Overall there is a clear preference for microstates in the ′bound′ macrostate to show lower distances between the ligand and the peptide than in the ′unbound′ macrostate, however some overlap remains, which could be due to the broad range of MD snapshots assigned to a single microstate.

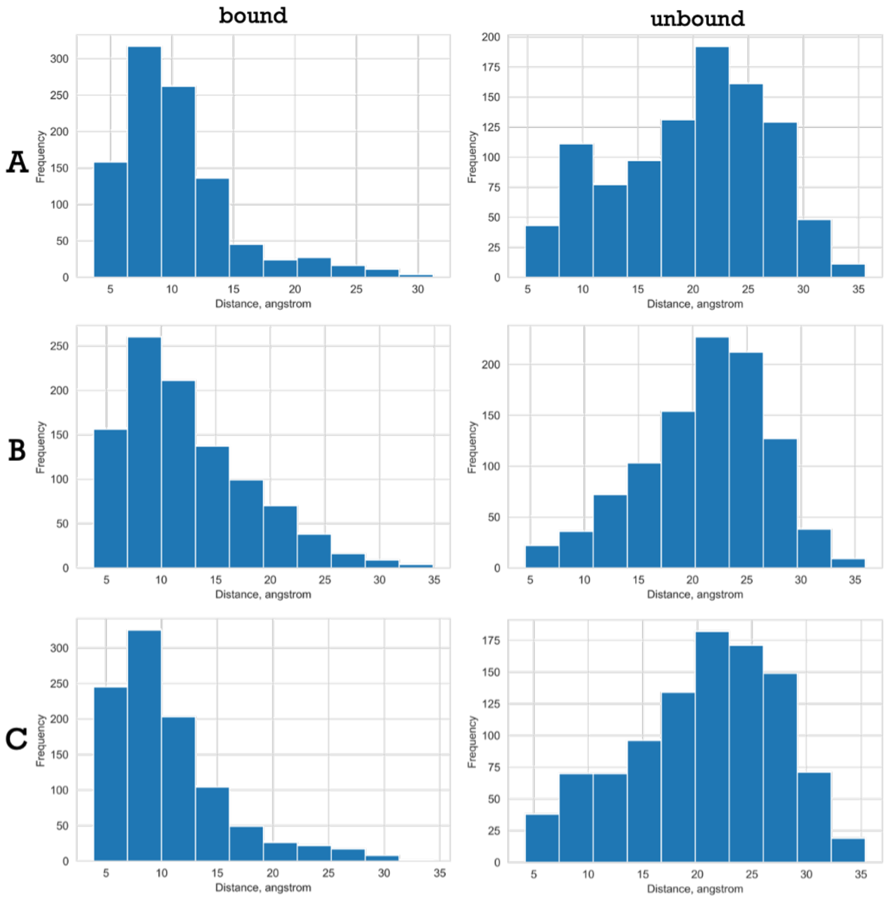


FIGURE 4.13: Probability distribution of distances of the bound and unbound states of the second metric applied for the three force fields, **A** AmberIDP, **B** Charmm36m, **C** FF14SB.

The stationary probabilities of these states, π_1 for the unbound state and π_2 for the bound state, were computed by summing over all the microstates and are reported in the following table:

SOMD/ff14SB/GAFF2		
Metrics	π_1	π_2
Metric 1	0.75 ±0.05	0.25 ±0.06
Metric 2	0.48 ±0.05	0.52 ±0.06
Metric 3	0.47 ±0.06	0.53 ±0.09
GROMACS/Charmm36m/Charmm		
Metric 1	0.72 ±0.02	0.28 ±0.04
Metric 2	0.47 ±0.05	0.53 ±0.03
Metric 3	0.51 ±0.02	0.49 ±0.04
GROMACS/Amberidp/GAFF2		
Metric 1	0.47 ±0.02	0.52 ±0.05
Metric 2	0.32 ±0.02	0.68 ±0.08
Metric 3	0.32 ±0.03	0.68 ±0.01

TABLE 4.2: Stationary probabilities of the bound (π_2) and unbound (π_1) states for the three metrics of each force-field.

The values of the stationary distributions highlighted that the bound state was more dominant for metrics 2 and 1 but it differed in population amongst the three force-fields. On the other hand, metric 1 showed that the unbound state was the most favorable for ff14SB and Charmm36m. It was apparent that the Amberidp force-field showed the strongest tendency for 10058-F4 to bind favorably to the c-Myc$_{402-412}$ peptide.

In order to provide a more quantitative interpretation of this binding process, we calculated both the binding affinity and the kinetics of the overall process of binding for the three metrics for each force-field. The overall binding free energy was computed from the stationary distributions and the standard state correction term as it was described in the Methods section. The values obtained from this calculation for the first two metrics were similar between the three force-fields, and were only 1 kcal/mol more negative than those ones obtained from the absoluteFEP protocol. However, for the case of metric 1, the binding free energies computed for Charmm and FF14SB were similar with the binding free energy obtained from the absoluteFEP protocol, when custom parameterisation was used, and differed from the corresponding binding free energy of the Amberidp force-field.

This finding highlights the reproducibility amongst the methods, although the details of the simulated conformational ensembles differ.

SOMD/ff14SB/GAFF2				
Metrics	Average distance	V$_{bound}$	V$_{unbound}$	$\Delta G^{\circ}_{\mathrm{msm}}$
Metric 1	10 ± 4	4100 ± 1300	75000 ± 1300	-1.60 ± 0.66
Metric 2	10 ± 5	5000 ± 2000	74100 ± 2000	-2.30 ± 0.71
Metric 3	11 ± 5	5000 ± 1700	74100 ± 1700	-2.32 ± 0.68
GROMACS/Charmm36m/Charmm				
Metric 1	8 ± 2	2600 ± 540	76000 ± 600	-1.62 ± 0.77
Metric 2	12 ± 5	8100 ± 540	70000 ± 2700	-2.28 ± 0.51
Metric 3	12 ± 5	7700 ± 2500	71000 ± 2500	-2.21 ± 0.77
GROMACS/Amberidp/GAFF2				
Metric 1	8 ± 2	2600 ± 520	77000 ± 600	-2.33 ± 0.48
Metric 2	10 ± 4	4700 ± 1500	75000 ± 1500	-2.71 ± 0.66
Metric 3	10 ± 4	5000 ± 1600	74000 ± 1600	-2.71 ± 0.68

TABLE 4.3: Average distance between the com of the ligand and the com of the protein in the bound macrostate are given in Å, volumes of the bound and unbound states are given in Å^3, standard binding free energies are given in kcal/mol.

Another important feature that can be computed from the MSM models is the kinetics that govern the binding process of 10058-F4 to c-Myc. For this purpose Mean First Passage Time values (MFPTs) between the two states in each force field for the three metrics were calculated from the Bayesian MSM using the pyEMMA software. MFPT values were then converted into rate constants for the binding and unbinding of 10058-F4 to c-Myc (using a concentration of the ligand equal to 0.02 M given the box dimensions). The k$_{on}$ and the k$_{off}$ values for each of the three force-fields for the three metrics are summarised in the following table:

SOMD/ff14SB/GAFF2				
Metrics	MFTP$_{1->2}$	MFTP$_{2->1}$	k$_{on}$	k$_{off}$
Metric 1	14.28 ± 0.04	15.03 ± 0.04	317.04 ± 0.83	7.00 ± 0.02
Metric 2	14.76 ± 0.04	15.88 ± 0.04	300.08 ± 0.79	6.78 ± 0.02
Metric 3	28.80 ± 0.20	9.30 ± 0.15	512.29 ± 3.65	3.47 ± 0.05
GROMACS/Charmm36m/Charmm				
Metric 1	11.32 ± 0.02	12.62 ± 0.03	373.69 ± 0.79	8.84 ± 0.01
Metric 2	13.84 ± 0.03	13.95 ± 0.03	340.70 ± 0.67	7.17 ± 0.01
Metric 3	7.37 ± 0.04	22.25 ± 0.10	212.02 ± 0.10	13.56 ± 0.07
GROMACS/Amberidp/GAFF2				
Metric 1	13.10 ± 0.02	28.93 ± 0.08	165.39 ± 0.39	7.63 ± 0.02
Metric 2	15.84 ± 0.04	35.11 ± 0.12	136.23 ± 0.33	6.31 ± 0.02
Metric 3	16.59 ± 0.08	14.98 ± 0.13	319.36 ± 2.73	6.03 ± 0.05

TABLE 4.4: Mean first passage times (MFTPs) between the unbound (1) and the bound (2) state as were estimated from the Bayesian MSM. MFTPs are measured in ns. The kinetic reaction rates of the three different force-fields for the three metrics, k$_{on}$ and k$_{off}$ for the bound and the unbound states respectively. k$_{off}$ is reported in μs^{-1} and k$_{on}$ in μs^{-1} M^{-1}.

The computed k$_{on}$ and k$_{off}$ values from our MSM models can be compared with values that would be expected for a protein-ligand binding process. The typical range of the k$_{on}$ rates spans between 10^{3} s^{-1} M^{-1} to 10^{9} s^{-1} M^{-1}, with the latter corresponding to the rate limit of diffusion of a solute to the solvent.[302] Thus the on rate constants computed (ca. 10^{8} s^{-1} M^{-1}) from the MSMs are close to diffusion limit. k$_{off}$ values typically range from 1 s^{-1} to around 10^{7} s^{-1}, due to the long-lasting nature of protein-ligand interactions.[302] The MSM-derived k$_{off}$ values appear therefore to be at the upper range of what is experimentally observed. Overall the picture that emerges is one of weak affinity with very fast binding/unbinding kinetics.

A final aim of this study was to identify the residues that the ligand prefers to interact with when in its bound state. For this purpose, 1000 snapshots were extracted based on microstates probabilities using the appropriate pyEMMA functions in order to create a trajectory with snapshots from the bound state of metric 2. Metric 2 showed the strongest tendency for 10058-F4 to bind favorably to the c-Myc$_{402-412}$ peptide for all the applied force-fields. In the resulting trajectory, we applied a homemade python script using the *mdtraj* module to count the number of carbon atoms of the ligand and measure the distance of those atoms with the carbon atoms of the protein. This metric allowed us to evaluate the hydrophobic contacts of the ligand with every residue for each snapshot. A cutoff of 4Å was used for every distance to only identify close contacts between 10058-F4 and each residue of the oncoprotein c-Myc. We also examined the ability of the ligand to engage in hydrogen bonding interactions with each residue using the cpptraj module. The total number of hydrophobic contacts and hydrogen bonds formed during the 1000 snapshots for Amberidp force-field is depicted in the Figure 4.14, while the total number of contacts for the other two force-fields are given in the Appendix.

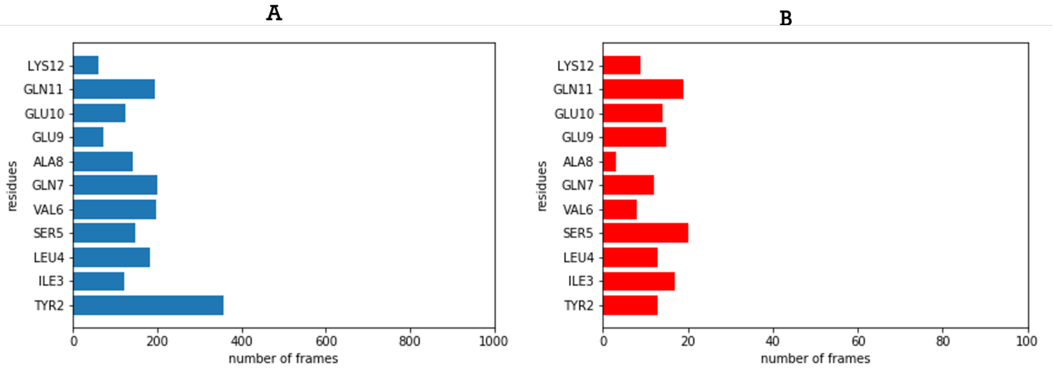


FIGURE 4.14: A) Total number of hydrophobic contacts (on the left with blue color) and B) Number of hydrogen bonds (on the right with red color) for Amberidp force-field. The x-axis has been truncated to 100 for visualisation purposes (number of contacts less thab 100.

The results for each force-field indicated that the ligand prefers to bind in the N terminus of c-Myc$_{402-412}$ especially with Tyr402. In addition, the binding process is mainly characterised by van der Vaals interactions rather than hydrogen bonds, since only a small fraction of the 1000 snapshots involve hydrogen bonding interactions between the ligand the peptide. Five representative snapshots for the Amberidp-Metric 2 analysis were obtained from the corresponding trajectory and are depicted in the following figure.

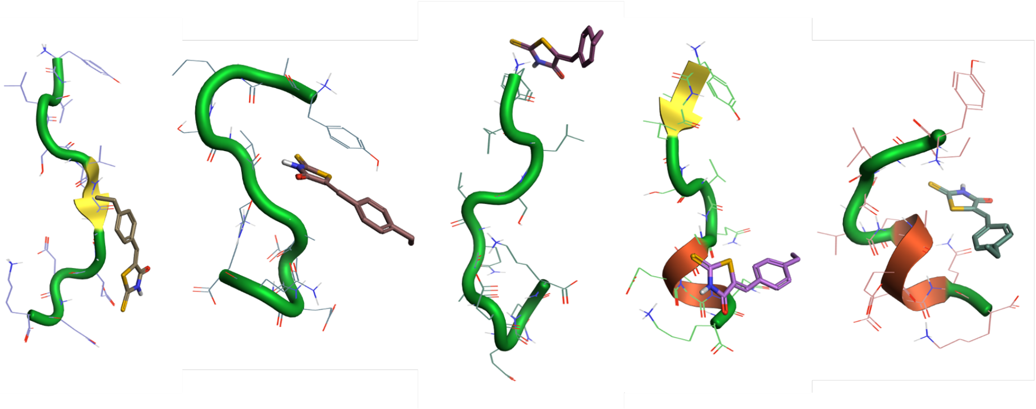


FIGURE 4.15: Five representative snapshots from the bound state of metric 2 for the Amberidp force-field.

These snapshots highlight the flexibility of the bound state, as c-Myc and 10058-F4 adopt diverse conformations. Finally, we observe that the nature of the bound state described by the MSM appear to be in line with the findings from the previous metadynamics study of Michel and Cuchillo.[290]

4.4 Conclusions

Two molecular dynamics simulation protocols were established to study the interactions of small molecules with the intrinsically disordered protein c-Myc. Alchemical free energy calculations were first applied to compute the absolute binding free energies of c-Myc$_{402-412}$/10058-F4 and c-Myc$_{402-412}$/10074-G4 complexes. This protocol was developed in Chapter 3 and it generated reproducible results for this system. However the computed free energies of binding (ca. -2 kcal.mol^{-1}) deviated significantly from experimental data (ca. -6 kcal.mol^{-1}).

The second protocol was employed in order to understand the binding process of the small molecule 10058-F4 to c-Myc. For this purpose, extensive MD simulations and Markov State models were combined to reversibly simulate binding/unbinding in the c-Myc$_{402-412}$/10058-F4 complex. The binding/unbinding kinetics of the protein-ligand complex can be described as a two-state process because the slowest transitions are due to the binding process. We discretised the MD trajectories into bound and unbound states by using as features the distances between some parts of the ligands and the CA atoms of the c-Myc residues.

The binding free energies obtained from these models were broadly similar to those obtained from the absoluteFEP protocol. These result highlight difficulties for the present molecular modelling protocols to reproduce experimental trends for this system. For this purpose, this study also focused on the residues that critically affect binding as well as the type of the interactions that govern the binding process. The protein amino acids that were located in the N-terminus of c-Myc$_{402-412}$ were more crucial for binding and the interactions between them and the ligand were mainly hydrophobic. This is corroborated by a study implemented in 2012 from our group using a bias-exchange variant of the metadynamics method (BEMD) to sample extensively the energy landscape of c-Myc$_{402-412}$/10058-F4 complex. This study also highlighted these residues as crucial for binding.[290]

To test the accuracy of our method we computed rate constants between the two states and compare them with the expected values for k$_{on}$ and k$_{off}$ based on experimental studies.[302] Our predicted values are at the upper ranges of what may be observed experimentally, suggesting the models describe a 'fast and weak' binding scenario.

Finally, the use of three different force-fields for the parameterisation of c-Myc also provided insight into the modelling of IDPs. The stationary probabilities of the two states were somewhat sensitive to the choice of the force-field as the bound states were more favored by using the f14IDPSFF Amber parameters for the oncoprotein.[293] However none of the variability observed translates into significantly different standard binding free energies. Similarly customised GAff parameters for 10058-F4 did not affect dramatically the binding free energy obtained from the absoluteFEP protocol.

Overall, we demonstrate the performance of two different MD based protocols in computing binding free energies of c-Myc/ligand complexes and characterising the binding mechanism of a small molecule to c-Myc.

The results are consistent between the two methods but are in disagreement with the experimental findings. A possible reason for this inconsistency could be that the mechanism of binding of the ligand to c-Myc is more complex than the 1:1 stoechiometry assumed here. However, despite the difficulties needed to overcome, our MSM models were effective to define binding as a two-state process and unravel the potential binding site and types of interactions between c-Myc$_{402-412}$ and 10058-F4. Our work suggests that more efforts should be directed to predict via computational methods whether a given small molecule will bind to an IDP such as c-Myc. Our models can be used as a starting point for future work that could consider other force fields.

Chapter 5

Absolute Binding Free Energy Calculations of Ligands for the Flexible Protein MDM2

5.1 Introduction to MDM2

5.1.1 General information on MDM2/p53 interaction

Over the last four decades the protein p53 has been recognised as a crucial transcription factor for the protection of the integrity of the genome.[303] p53 is encoded by the TP53 gene and consists of two N-terminal transactivation domains (TADs) followed by a proline rich domain, a DNA binding domain, a C-terminal domain that encodes nuclear localization signals, and an oligomerization domain required for transcriptional activity. Upon activation through various stress signals, p53 can efficiently inhibit the proliferation of precarcinogenic and carcinogenic cells by both blocking cell cycle progression and inducing apoptosis.[304–306] In addition to these functions, p53 can prevent cancer development through non-canonical p53 activities such as regulation of microRNA processing, anti-oxidant response, modulation of tumor stroma and immune response and regulation of metabolism and autophagy.[307]

Inactivation of p53 function is needed for the development and maintenance of a variety of tumors.[308] Impairment of p53 tumor suppression function occurs through two general mechanisms. First, point mutations in p53 inactivate its tumor suppressor function, a situation observed in nearly 50% of human cancers.[309] The most well-characterised and common TP53 mutations occur in the region encoding p53's DNA binding domain, indicating that this region is crucial for preventing cancer development. Reactivation of p53 function can be achieved with small molecules that bind to full-length p53[310, 311] or the core DNA-binding domain of mutant p53.[312] This strategy has resulted to the development of a potent small molecule, PRIMA-1MET/APR246, that has progressed to Phase III clinical trials.[313]

In the remaining 50% of human cancers, wild-type p53 is maintained at low levels by a variety of mechanisms. One major inhibitory mechanism of p53's transcriptional activity is through overexpression of murine double minute-2 (MDM2) protein.[305] MDM2 and its homologue MDMX are the two primary negative regulators of p53. MDM2 regulates the cellular levels of p53 via different mechanisms. Firstly MDM2 recognises the p53 TAD domain with its N-terminal, domain and the formation of the p53/MDM2 complex blocks its transcriptional activity.[314] Secondly MDM2 is an E3 ubiquitin ligase that can either monoubiquitinate p53, preventing the binding of p53 to DNA and promoting exportation of p53 out of the nucleus, or polyubiquitinate, promoting proteasomal degradation of p53.[315] MDMX does not have an intrinsic E3 ligase activity, but it does inhibit p53's transcriptional activity by formation of inactive complexes. MDMX can further form a heterocomplex with MDM2 that enhances MDM2's E3 ligase activity.[316, 317] MDM2 and MDMX participate thus in an important feedback loop that is illustrated in Figure 5.1.

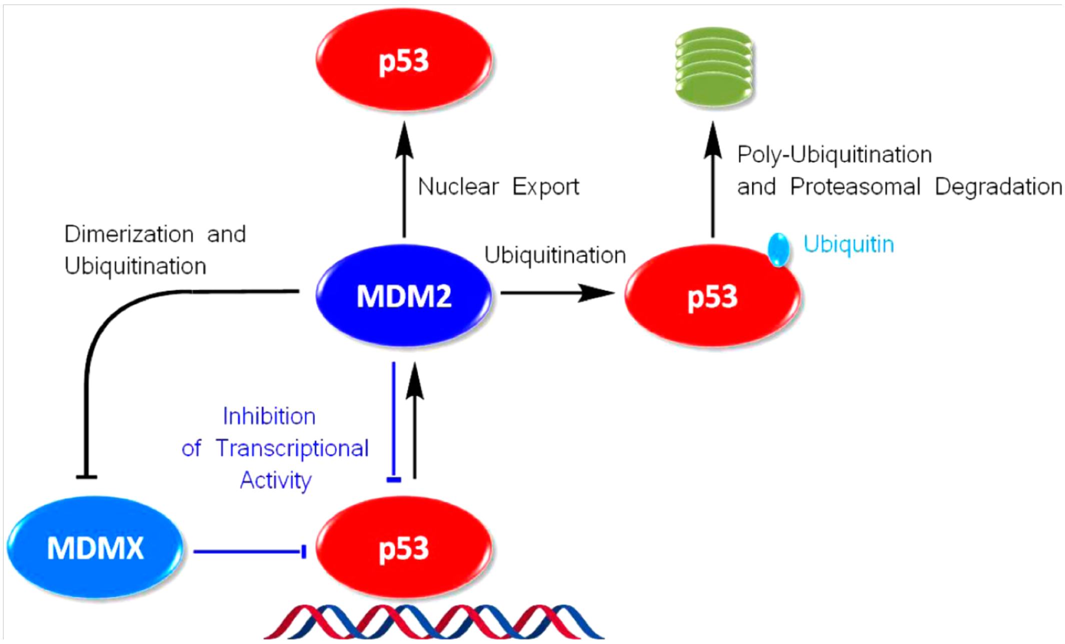


FIGURE 5.1: Depiction of autoregulatory feedback loop of MDM2/p53. p53 transcribes MDM2 and increases MDM2 expression. MDM2 inhibits p53 tumor suppression function via three different mechanisms: 1) ubiquitination of p53 that leads to proteosomal degradation, 2) export of p53 out of the cell nucleus and 3) reduction of p53 transcriptional activity through binding to p53 TAD domain. Adapted from Wade *et al.*[317]

Due to the importance of MDM2 and p53 in cancer, a significant body of work has focused on blocking the MDM2/p53 interaction to restore the transcriptional activity of wild-type p53. The p53 TAD domain adopts an α-helical conformation during its interaction with MDM2 through three hydrophobic residues, Phe19, Trp23, and Leu26, that protrude into three hydrophobic pockets of MDM2.[318, 319] Because the "Phe19-Trp23-Leu26 hydrophobic cleft" is compact and well-defined it is feasible to identify small molecules that block formation of the MDM2/p53 complex via binding to MDM2.

Thus a range of small molecules have been developed to bind strongly to MDM2. The first reported MDM2 inhibitors with *in vivo* activity contain a cis-diphenyl substituted imidazoline scaffold and are also called nutlins.[320] Hoffmann-La Roche group designed Nutlin-3a, the first successful example of a molecule that mimics the molecular interactions between MDM2 and p53 and binds to MDM2 with IC_{50} = 90 nM. Further optimisation of Nutlin-3a led to the RG7112 analogue, the first MDM2 inhibitor to advance into clinical trials for the treatment of liposarcoma patients with MDM2 amplification.[321] Finally, RG7388 binds to MDM2 with IC_{50} = 6 nM and is more potent in induction of p53 activation *in vivo* than other nutlins.[322]

Other remarkable examples are the spirooxindole-containing compounds (MI series) from the University of Michigan with IC_{50} values of 30–2000 nM.[323] MI-888 had an excellent MDM2 binding affinity and showed antitumor activity without obvious adverse effects upon oral administration.[324] This study has led to the discovery of MI-773 (SAR405838) that was advanced in phase I clinical trials by Sanofi in 2012.[325] Benzodiazepene-based derivatives from Johnson & Johnson also provided evidence of high affinity (IC_{50} values of 0.5–2 μM) and suppression of the growth of cell lines containing wild-type p53.[326] Finally, a piperidinone class of MDM2 ligands were discovered by Amgen.[327] A structural analysis of previously known MDM2 inhibitors yielded compound AM-8553 that had an IC_{50} value of 1.1 nM to MDM2 and also showed a dose-dependently anti-tumor effect in SJSA-1 osteosarcoma xenograft mouse model.[328] This finding led to a series of potent and orally active MDM2 inhibitors including AMG-232, that was selected for clinical trials in patients with different types of solid and hematological tumors.[327] Finally, replacement of the carboxylic acid of AMG-232 with a 4-amidobenzoic acid resulted to AM-7209, the most potent (K_D = 38 pM from ITC measurements) and selective inhibitor from this class of compounds with improved pharmakocinetic properties. It has remarkable in vivo antitumor activity in both HCT-116 colorectal carcinoma xenograft model and the SJSA-1 osteosarcoma xenograft model and it is considered as the most promising molecule for the treatment of human cancer.[328] The chemical structures of the aforementioned MDM2 inhibitors are illustrated in Figure 5.2.

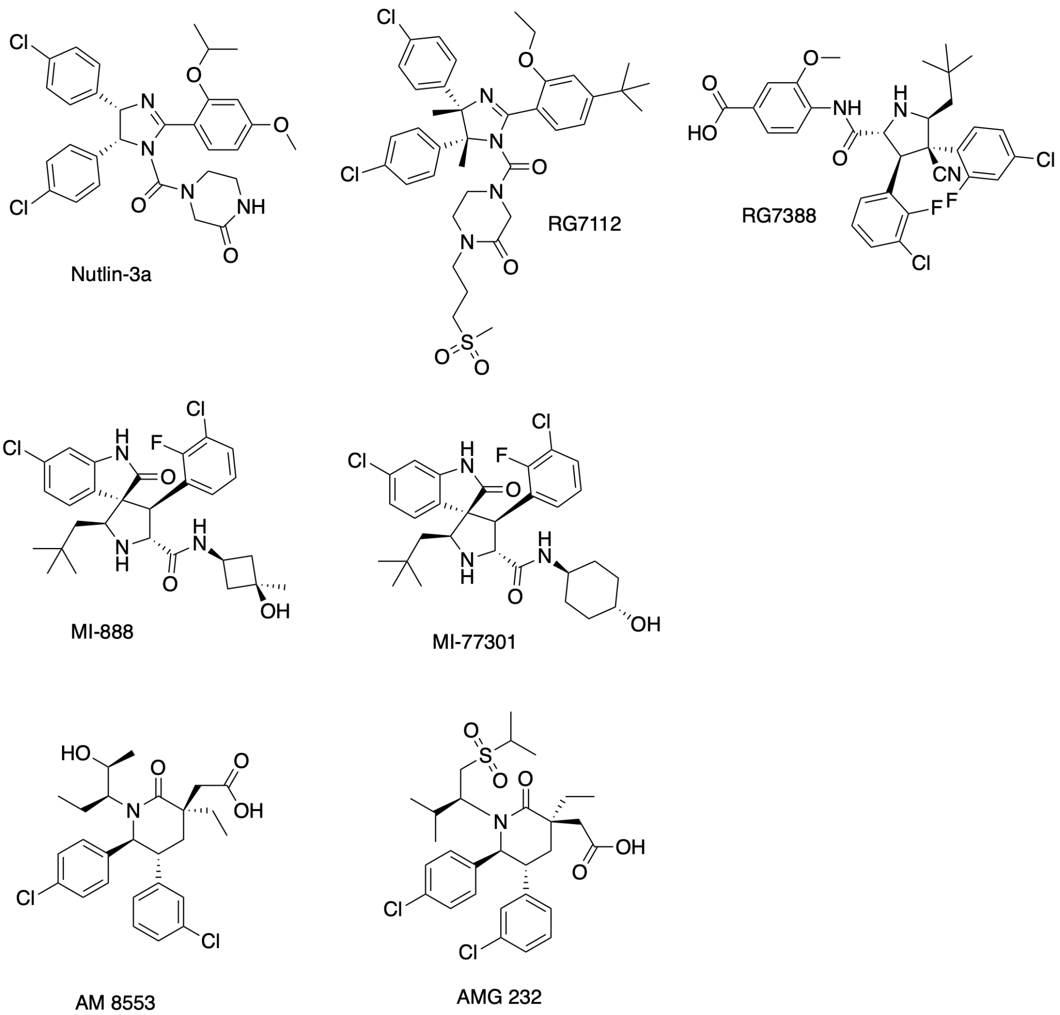


FIGURE 5.2: Chemical structures of small molecule inhibitors that target MDM2/p53 interaction.

Most of these drug discovery efforts have focused their attention on residues 25-120 of MDM2 that are natively structured. This region is called MDM2 core region and contains the "Phe19-Trp23-Leu26 hydrophobic cleft". However, the first 24 residues of the N terminal domain of MDM2 form a flexible lid that is an intrinsically disordered region in native conditions and can adopt both "open" or "closed" states(Figure 5.3). In the "closed" state, the lid competes with p53 for binding to the p53 binding site of the core region via a pseudo-substrate mechanism.[329] In addition, Showalter *et al* used NMR experiments to show that the fluctuation between the two states occur on a time-scale greater than 10 ms.[330]

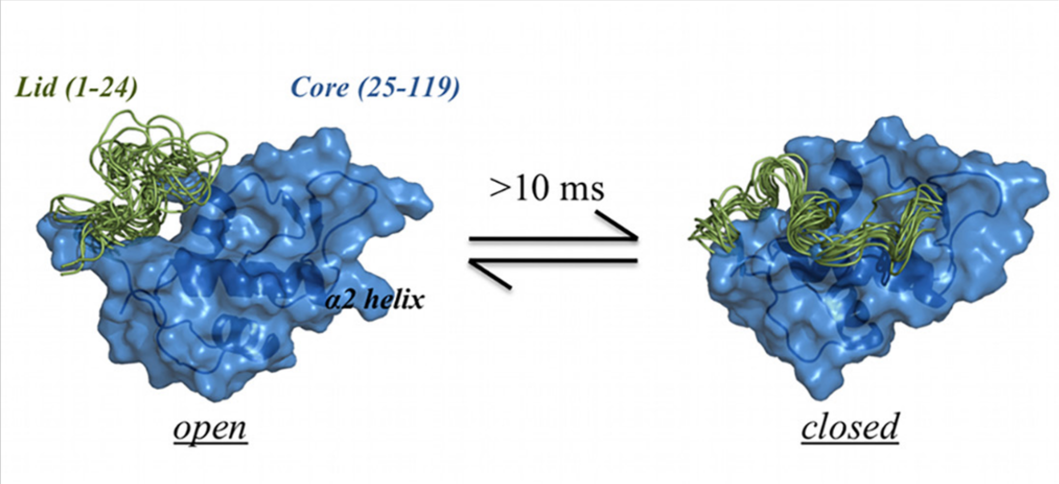


FIGURE 5.3: The interconversion of the lid (residues 1-24 in green) between "open" and "closed" state occurs in ms time scale. In the latter state, the lid occupies the "Phe19-Trp23-Leu26 hydrophobic cleft" in the MDM2 core region (residues 25-120 in blue). Adapted from Bueren-Calabuig *et al.*[331]

The lid region has not historically been considered as a “hot spot” for the design of potent small molecule inhibitors by structure-based drug design campaigns due to its high structural plasticity. However in 2013, Michelsen *et al.* showed that the piperidinone class of MDM2 inhibitors establish interactions with the lid region that are crucial for MDM2 recognition.[332] Such compounds are able to order the lid region on binding, in spite of the entropic cost incurred. For instance, a piperidinone inhibitor (Pip-2) showed a 25-fold preference in binding with a long MDM2 construct (1-125) rather than with a short MDM2 construct (aa. 17-125). Such variability in binding preference is not observed for in the case of p53 peptide and Nutlin-3a binding.

5.1.2 Previous work from the Michel group and scope for further improvements

Bueren-Calabuig and Michel have previously used molecular simulations to progress understanding of the structure-activity relationships that enable the Piperidinone ligands to induce this remarkable conformational change of the MDM2 lid region. Specifically they employed accelerated molecular dynamics (aMD), umbrella sampling (US), and variational free energy profile (vFEP) methods to build atomically detailed lid structural ensembles between the flexible lid region and four ligands (p53, Nutlin-3a, Pip-2 and 1,4-benzodiazepine-2,5-dione (Bzd)). Analysis of the resulting lid structural ensembles (Figure 5.4) showed indeed that the flexible lid exhibits different conformational preferences with different classes of small-molecules. Pip-2 is the only ligand that induces the formation of a α-helical/β-strand structure due to hydrophobic contacts between Pip-2 and the lid. Nutlin-3a and Bzd compounds also show similar affinity for this structure, but their solubilising groups hinder its formation owing to interactions with segments of the lid that are more disordered. Therefore, this study highlighted the importance of the MDM2 lid for rational drug design of potent small-molecule inhibitors.[331]

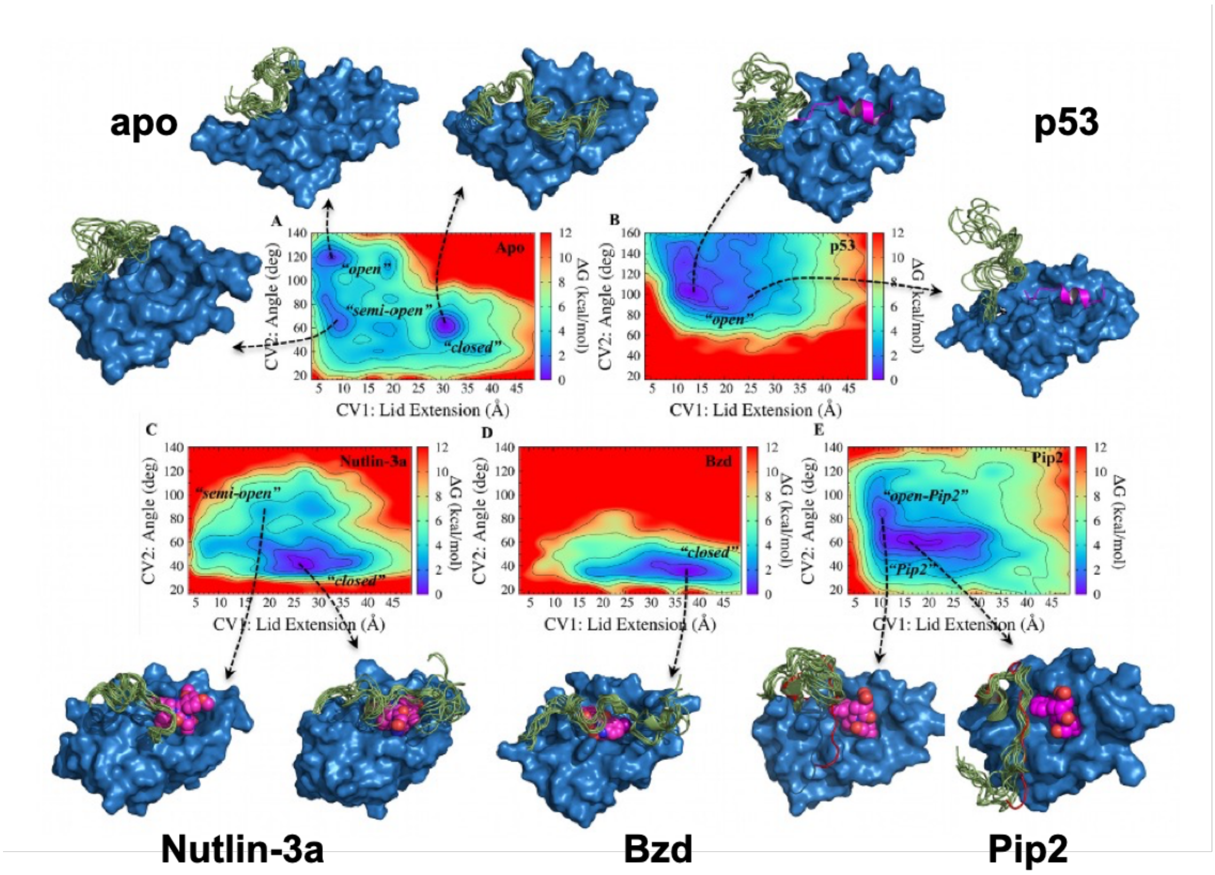


FIGURE 5.4: The computed Free Energy Surfaces (kcal mol^{-1}) from the resulting lid structural ensembles illustrated as a color coded heat map. Free energies are relative to the lowest free energy bin and are shown up to 12 kcal mol^{-1} above the lowest free energy bin. Ten representative lid conformations are shown for A) apo MDM2. B) p53(17–29)/MDM2 C) Nutlin-3a/MDM2 D) Bzd/MDM2 E) Pip-2/MDM2. Adapted from Bueren-Calabuig *et al.*[331]

The aim of this project is to build on the computational study of Bueren-Calabuig *et al.* to reproduce binding selectivity trends between two variants of MDM2 and five known MDM2 inhibitors.[331] Three of these inhibitors belong to the piperidinones class of MDM2 ligands that are able to structure the disordered lid region, while the other two compounds belong to the nutlins that do not demonstrate this ability. The two different MDM2 constructs differ in the length of the lid region. In the short construct the lid is truncated. The long construct the lid is largely present and can adopt two distinct conformational states. The first state is a structured and 'ordered' lid conformation similar to that observed when Piperidones bind to MDM2. The second state is an unstructured and extended lid state that wraps above the ligands, as suggested by the simulation study of Bueren-Calabuig and Michel. The chemical structures of the five MDM2 ligands together with the three MDM2 conformations considered are depicted in Figure 5.5.

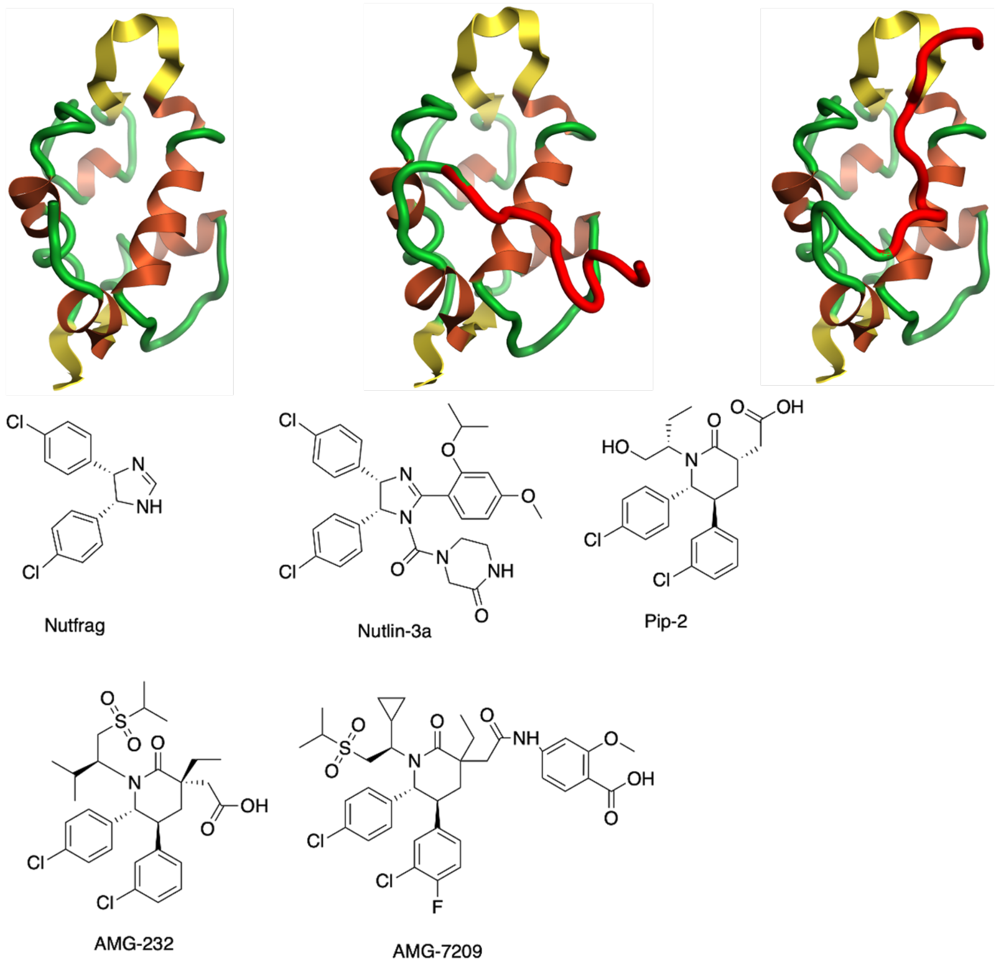


FIGURE 5.5: The three MDM2 conformations together with the chemical structures of the small molecule inhibitors that will be evaluated for their binding selectivity trends.

The standard binding free energies of the MDM2/ligand complexes will be computed through an absolute binding FEP protocol similar to that used in Chapter 3. However, as was observed in Chapter 4, accounting for the flexibility of intrinsically disordered regions is challenging for MD simulations. Further, the potein-ligand complexes considered here are structurally more complex than the host-guests studied in Chapter 3. Thus we first validate a novel adaptive sampling protocol in order to reduce the computational cost for this task. We also expand the distance-restraints framework used in Chapter 3 to facilitate convergence of the standard free energies of binding. The overall predictive power of the protocol is bench-marked against measurements of binding affinities collected using calorimetry experiments.

5.2 Methods

5.2.1 Preparation of MDM2/ligands input files for free energy calculations

For this computational study ,the protein-ligand structures were taken from the X-ray crystal structures with PDB IDs: 4J74 for Nutfrag/MDM2, 4HG7 for Nutlin-3a/MDM2, 4HBM for Pip-2/MDM2, 4OAS for AMG-232/MDM2 and 4WT2 for AMG-7209/MDM2. All water molecules were removed from the structures and all proteins were capped at the C-terminal and N-terminal with N- methyl and acetyl groups respectively. The coordinates of the structured lid conformation for the five protein-ligand complexes was obtained from Pip-2/MDM2 crystal structure, while the coordinates of the "closed" lid conformation was taken from a molecular dynamics study performed by group member Dr. Salomé Llabrés.

Input files for the free energy simulations were created using tleap.[168] Protein parameterisation was performed using ff14SB Amber force-field[169], while ligands were parameterised using the GAFF2 forcefield [170],[171] and use AM1-BCC partial charges[172]. The system was solvated in a cubic box with TIP3P water molecules[173], with a minimum distance between the protein and the edges of the box of 12 Å. Counter ions were added to neutralize the total net charge. The same approach was followed for parameterising the ligand in the free phase.

Next an equilibration protocol was applied to relax the box size. Initially, energy minimization of the entire system was implemented with 1000 steps of steepest descent gradients, using sander.[168] Then, an NVT protocol was followed for 200 ps at 298 K, followed by an NPT equilibration for further 200 ps at 1 atm. Finally, a 2 ns NPT MD simulation was run with the SOMD software to reach a final density of about 1 g cm^{-3}.[221],[222] The final coordinate files were retrieved with cpptraj.

5.2.2 Adaptive sampling protocol

Alchemical free energy simulations were performed following the double decoupling protocol that was described in details in Chapter 3. For the MDM2-ligand complexes, both bound and free phase *discharging* step were run with twelve λ windows (0.000, 0.050, 0.100, 0.200, 0.300, 0.400, 0.500, 0.600, 0.700, 0.800, 0.900, 1.000), while 26 λ windows (0.000, 0.025, 0.050, 0.075, 0.100, 0.125, 0.150, 0.200, 0.250, 0.300, 0.350, 0.400, 0.450, 0.500, 0.550, 0.600, 0.650, 0.700, 0.750, 0.800, 0.850, 0.900, 0.950, 0.970, 0.990, 1.000) were employed for the *vanishing* step, both in complex and solvated phase.

The above protocol is computationally very expensive as it consists off 76 λ windows. In addition, we need to run long MD simulations for each window as we do not know *a priori* the sampling time required for the protein-ligand complexes to visit all their thermally accessible states and thus improve the convergence of the final free energy of binding. Therefore, there was a need to derive an improved protocol that could reduce the waste of simulation time to windows that do not play a major role to the overall uncertainty of the absolute binding free energy for each simulation. For this purpose, the novel adaptive sampling approach in this Chapter offers a promising solution as it allows us to conduct extravagant MD simulations to a small number of important λ windows.

This approach involves running an initial set of calculations in which each λ*window* is simulated for a duration of 5 ns with SOMD in the NPT ensemble. Temperature control is maintained with an Andersen Thermostat with a coupling constant of 10 ps^{-1}.[48], while pressure control is achieved by a Monte Carlo barostat that attempts isotropic box edge scaling every 100 fs. A 12 Å atom-based cutoff distance for the non-bonded interactions is used, using a Barker Watts reaction field with dielectric constant of 78.3.[223] Distances restraints are used to prevent the decoupled ligand from leaving the MDM2 binding site. The restraint protocol used for these simulations is described in the subsection below.

Free energy changes were estimated with the multistate Bennet acceptance ratio method as implemented in the Sire utility *analysefreenrg*.[50] To achieve a more robust estimation of free energies, each simulation was repeated five times, using different initial velocities drawn from the Maxwell-Boltzmann distribution and statistical uncertainties are reported as 95% of the standard error of the mean.

Following this, we identify a small number k«M of λ windows that contribute the most to the overall uncertainty in the absolute binding free energy. To do so, we calculate the standard deviation of the free energy estimates between neighboring MBAR windows $\Delta\Delta G_{i->j}$. We then tested different thresholds of the standard deviation of two windows and we concluded that if this threshold is higher than 0.1 kcal/mol, then these windows have a great impact at the overall uncertainty of our simulations. Once all the windows are selected, then additional replicates of these windows are simulated to improve estimates of the mean free energy change for these windows. Each round of simulations is denoted an epoch. In each epoch every λ window selected is simulated for a duration of 5 ns with SOMD in the NPT ensemble with the same protocol used above. This allows overall the calculation of precise binding free energies at a fraction of the computing cost (approximately six times) of the original protocol.

5.2.3 Restraints protocol

The choice of the restraints is important for these calculations as it influences the convergence of the final binding free energy. For instance, the ligands used to inhibit MDM2, due to their large size, can adopt multiple orientations during the course of the simulation once their interactions with the surrounding environment are weakened. This translates into an increase in computing time requirements in order to sample all the thermally accessible orientations and thus slow down convergence of the final free energy of binding. To reduce the sampling time needed for these calculations, the restraint protocol used in the bound phase of the alchemical free energy simulations in Chapter 3 was modified in order to prevent the ligand from tumbling and drifting away from the host cavity. For this purpose, a series of flat-bottom distance restraints were defined between four guest atoms and a different number of host atoms depending on the MDM2/ligand complexes.

For the piperidinone class of MDM2 ligands we used four host atoms. The corresponding residues for each long and short construct of MDM2 and the corresponding atoms for every piperidinone inhibitor are illustrated in Figure 5.6. The restraints parameters of eq. 3.1 were: D$_{ji}$ = 2 Å and k$_{ji}$= 10kcal mol^{-1} Å^{-2}. The R$_{ji}$ distance employed for each host-guest atom is depicted in Figure 5.6.

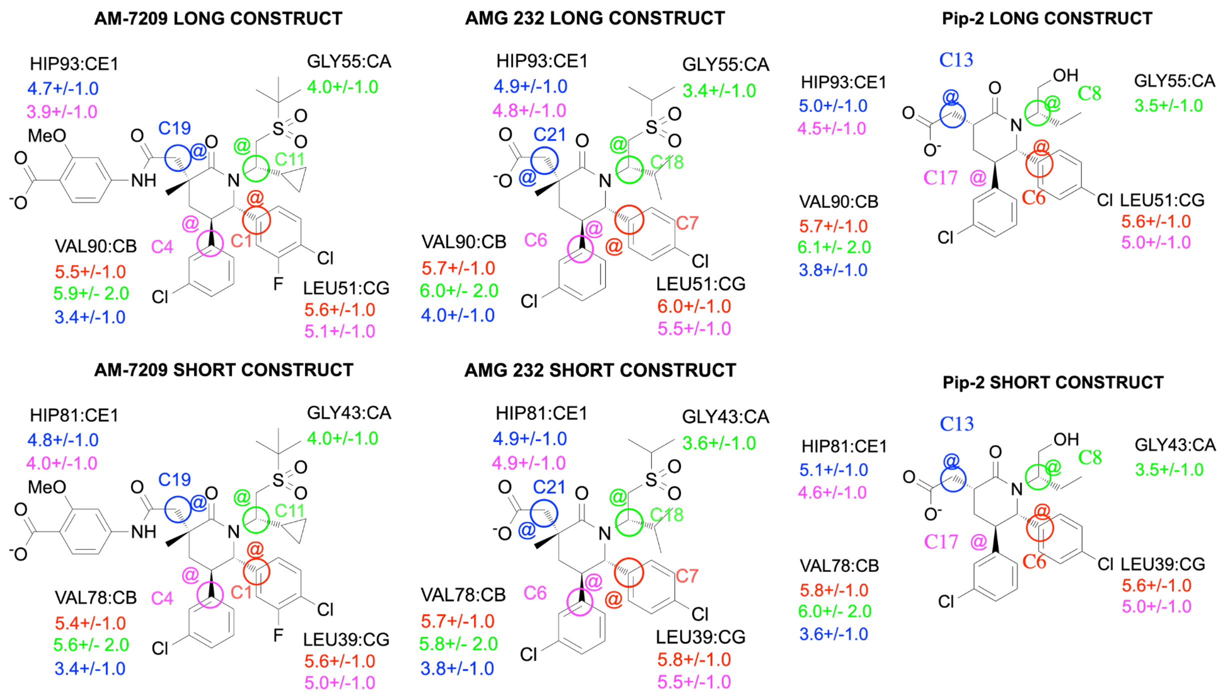


FIGURE 5.6: Illustration of the restraint protocol used for every MDM2/piperidinone complex. Each guest atom used for the flat-bottom distance restraints is depicted with a different color. The same color is used for their R_{ji} distances with the corresponding residues. Each host atom selected for the restraint scheme is characterised by the number and the name of the residue that it belongs.

For the nutlin inhibitors, six host atoms were used for Nutlin-3a and four atoms from MDM2 were selected for Nutfrag (Nut4). The restraints parameters of eq. 3.1 were: D_{ji} = 2 Å and k_{ji}= 10kcal mol^{-1} Å^{-2}. The R_{ji} distance applied for each host-guest atom together with the corresponding protein-ligand atoms used for the restraint protocol are showed in Figure 5.7.

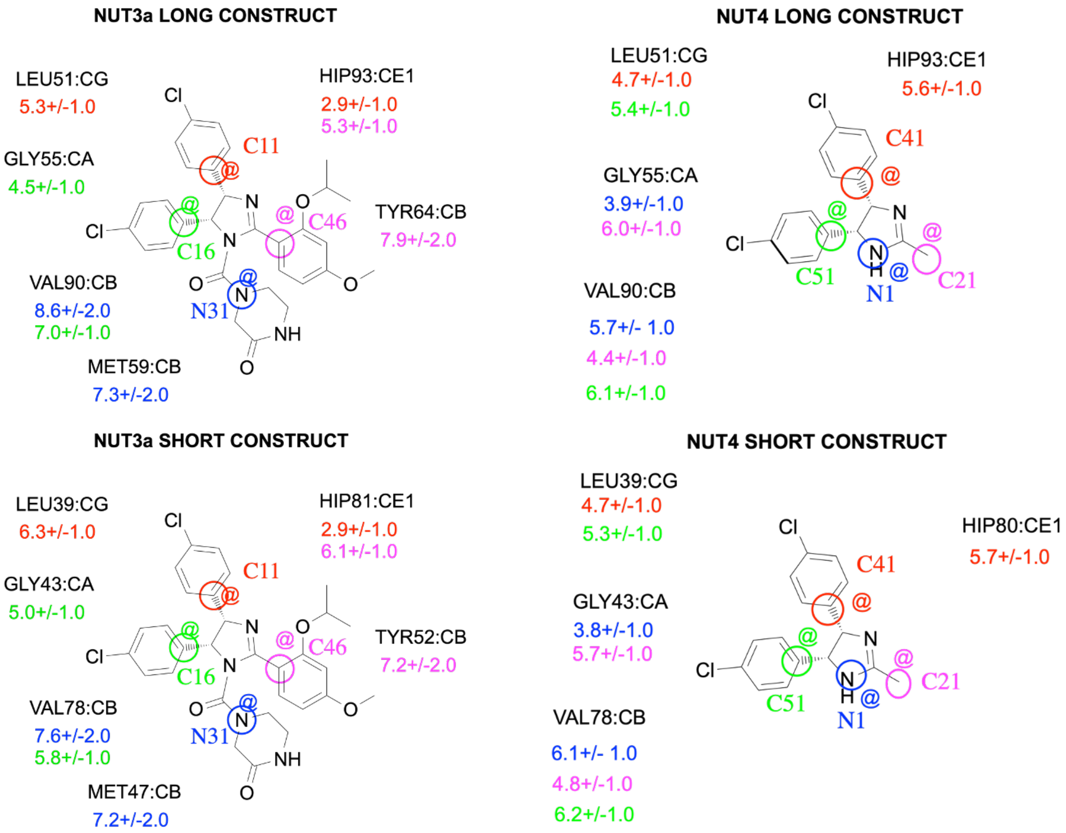


FIGURE 5.7: Depiction of the restraint protocol used for the MDM2/nutlin complexes. Each guest atom used for the restraint scheme is depicted with a different color. The same color is used for their R_{ji} distances with the corresponding amino acids. Each host atom selected for the flat-bottom distance restraints is characterised by the number and the name of the residue that it belongs.

5.3 Results

5.3.1 Choice of threshold for the adaptive sampling protocol

A critical aspect of the adaptive sampling protocol described in the previous section is that it is necessary to define a suitable threshold for the standard deviation of the free energy estimates between neighbouring λ windows. This threshold determines the number of λ windows that contribute the most to the overall uncertainty of the binding free energy. If the threshold is too high, the adaptive sampling protocol will terminate early, and may appear to converge to an answer that deviates from converged results. If the threshold is too low no significant time savings are achieved.

To better understand what a suitable value of the threshold parameter could be, we employed initially a brute force sampling protocol for the MDM2 short construct/Pip-2 complex. Each λ window was thus simulated for 50 ns (ten epochs) and each simulation was repeated five times using the protocol described in the Methods section. The binding free energies of the five independent runs were computed using equation 3.2. In addition, the average binding free energy of these simulations was also calculated. The graph of the convergence of these raw free energies of binding as a function of the cumulative sampling time (convergence plot) of the brute force sampling protocol is illustrated in the following figure:

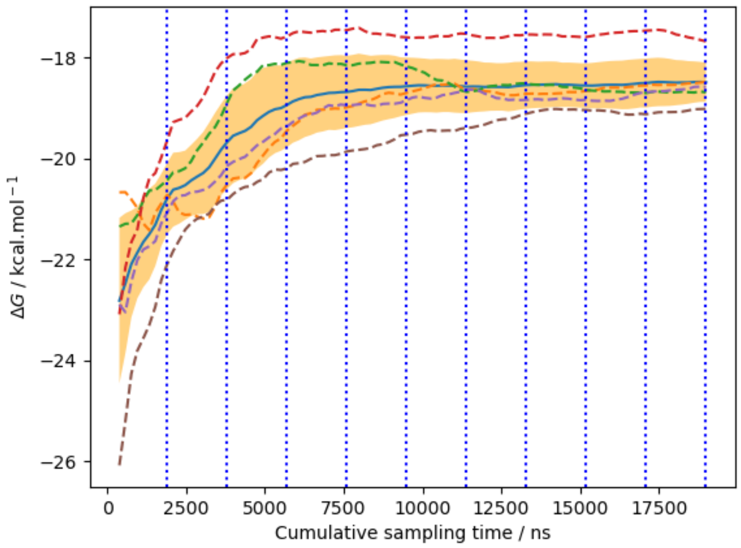


FIGURE 5.8: Depiction of the convergence plot of the MDM2 short construct/Pip-2 complex using the brute force sampling protocol. Dashed lines denote the binding free energies from the five different independent simulations started from different coordinates. The bold blue line denotes the average free energy of the five simulations, dotted blue lines represent each epoch and the shaded area denotes $\pm 0.95\sigma$.

Once the 50ns simulations were completed, we reanalysed the same dataset using variable thresholds (0.025 ; 0.050 ; 0.075 ; 0.100 ; 0.150 ; 0.200). For the first epoch, we calculated the binding free energies from the five different independent simulations using the same procedure as in the brute force sampling protocol. After the first epoch, we applied the different cutoffs to identify a subset of the λ windows that showed one standard deviation of the mean above the threshold. Then, we computed the binding free energies of the five independent runs using an additional 5 ns sampling time only for those selected λ windows. We iterated the same procedure for each protocol until we could not find neighbouring MBAR windows with standard deviation greater than the applied threshold, or until we reached the ten epochs of the brute force sampling protocol. The total number of windows that were selected for every alchemical step of the different protocols are illustrated in the windows sampling plots (Figures 5.10, 5.11). The calculated free energies of binding for each protocol at the end of the procedure are given in the table below:

Different protocols	
Protocols	$\Delta G_{average}$
brute force	-18.49 ± 0.39
Threshold 0.025	-18.13 ± 0.35
Threshold 0.050	-18.18 ± 0.43
Threshold 0.075	-18.09 ± 0.63
Threshold 0.100	-18.46 ± 0.57
Threshold 0.150	-19.45 ± 0.56
Threshold 0.200	-19.70 ± 0.67

TABLE 5.1: Results for all the protocols for MDM2 short construct/Pip-2 complex. Energies are reported in kcal/mol

The convergence plots produced from protocols using cutoffs 0.025, 0.050, 0.075, 0.100, 0.150, 0.200 are depicted in Figure 5.9.

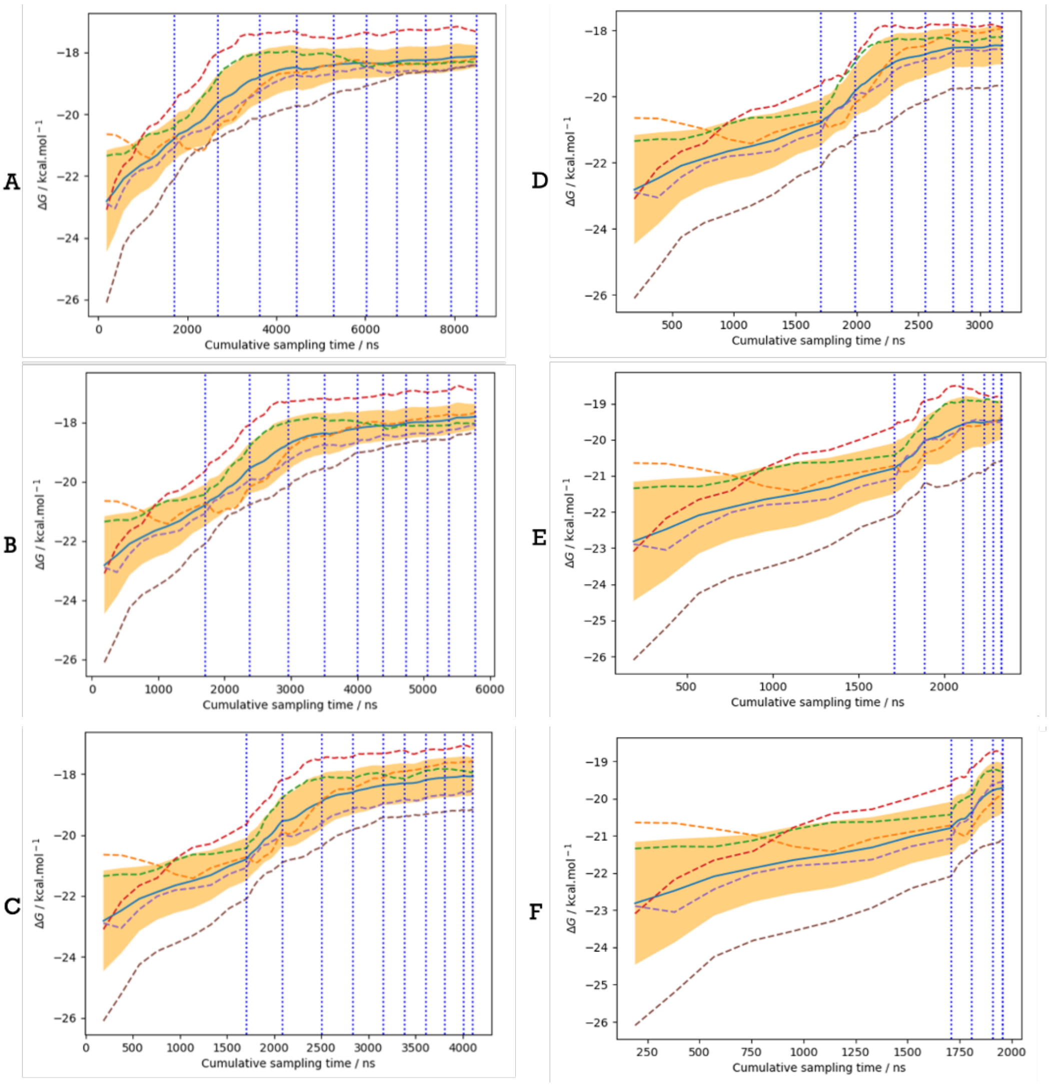


FIGURE 5.9: Convergence plots for protocols with cutoffs **A)** 0.025, **B)** 0.050, **C)** 0.075, **D)** 0.100, **E)** 0.150, **F)** 0.200.

The window sampling plots produced from protocols using thresholds 0.025, 0.050, 0.075, 0.100, 0.150 and 0.200 are shown in Figures 5.10 and 5.11.

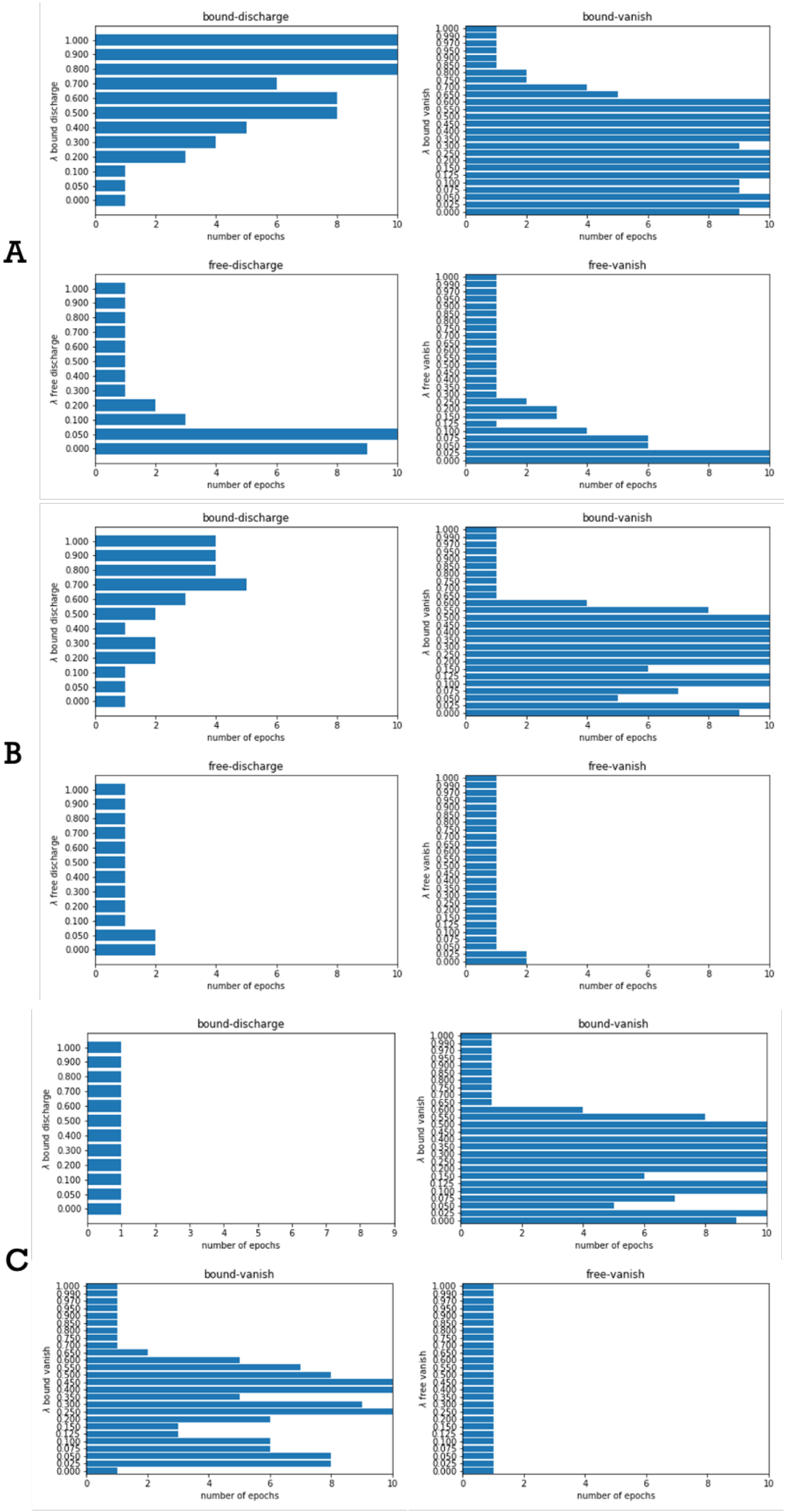


FIGURE 5.10: Convergence plots for protocols with cutoffs **A)** 0.025, **B)** 0.050, **C)** 0.075. Windows sampling plots consist off four subplots for each alchemical step of the absoluteFEP calculations. The name of the different number windows ranging from 0.000 to 1.000 is shown on the x-axis, while the number of epochs that each window was simulated is illustrated on the y-axis

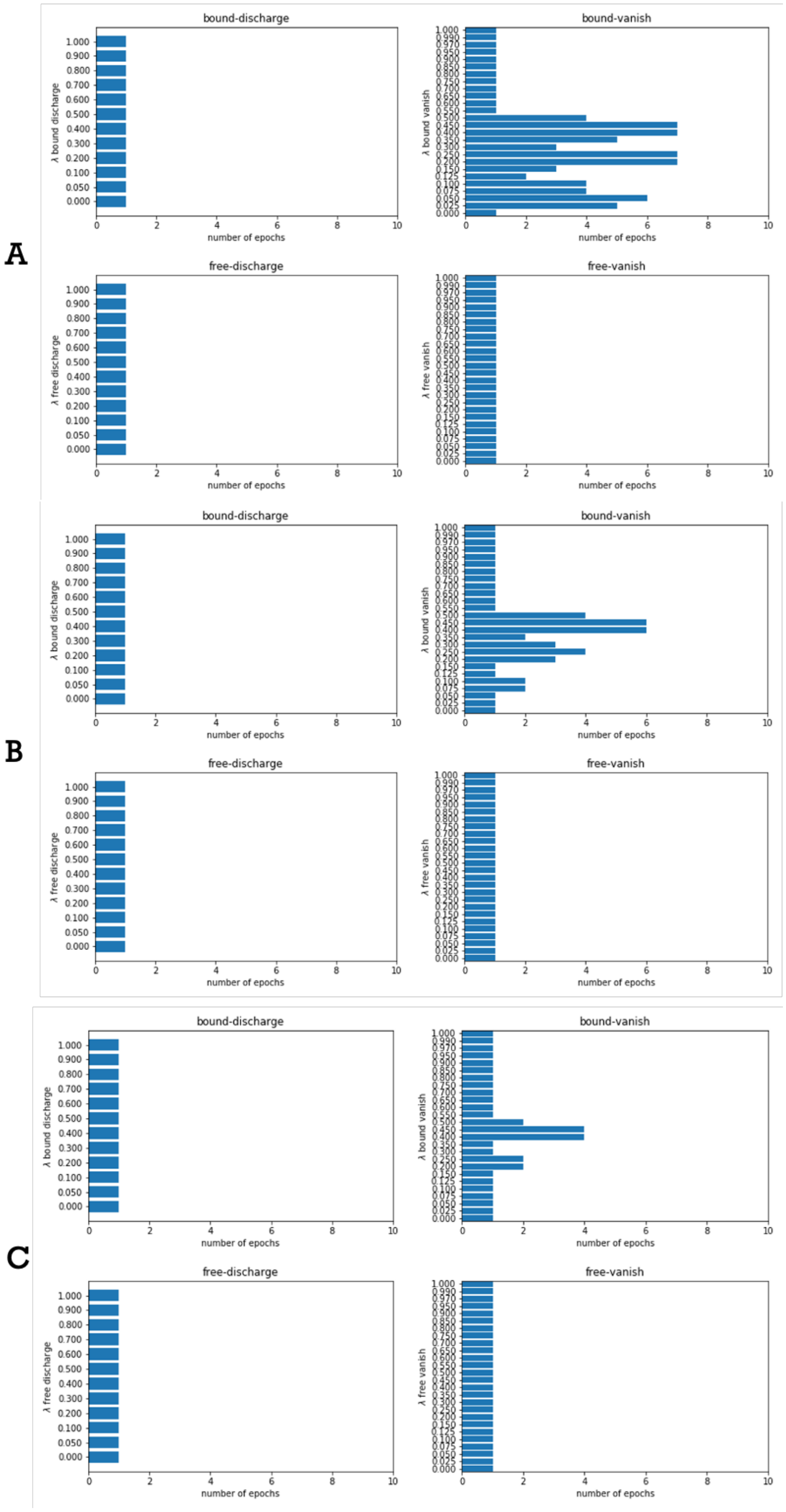


FIGURE 5.11: Window sampling plots for protocols with thresholds **A)** 0.100, **B)** 0.150, **C)** 0.200.

On the basis of the results, we observe that the alchemical steps that contribute the most to the uncertainty of the free energy estimates are mainly observed during the bound vanish stage of the double decoupling protocol. We observe that the adaptive sampling protocol with threshold = 0.100 kcal mol^{-1} gives us an estimated binding free energy statistically indistinguishable from the brute force results (-18.5 ± 0.6 kcal mol^{-1} versus -18.5 ± 0.4 kcal mol^{-1}), whilst achieving almost 6-fold decrease in computing requirements. Thresholds with lower values offer less savings in computing time, while thresholds with higher values result in binding free energy estimates that deviate from the full calculation owing to premature convergence of the adaptive sampling protocol. Therefore, we used the 0.100 kcal mol^{-1} threshold to process the rest of the dataset.

5.3.2 ITC results

A former member of Michel group, Dr. Cesar Mendoza Martinez, measured the free energy of binding of different MDM2/ligand complexes using ITC experiments. Titration data from these assays are provided in the Appendix. He used two MDM2 constructs, one with the disordered lid region present (residues 6-125) and one with the lid absent residues (residues 17-125). He tested all the ligands we used for our study, apart from Pip-2 for which we used the experimental data reported by Michelsen *et al.*[332] Measurements for MDM2-short/Nutfrag and MDM2-short/AMG-232 are less certain as only one or two replicates could be carried out for these two complexes, whereas measurements for all other complexes were performed in triplicates at least. The measured binding free energies, the change in enthalpy of the systems, ΔH and the temperature multiplied by the change in entropy of the systems, TΔS were also derived from the ITC experiments. All the aforementioned parameters are provided in the table below:

ITC measurements MDM2 short construct			
Ligands	ΔH	$-T\Delta S$	ΔG
Nutfrag	-5.01	-2.49	-7.50 ± 0.30
Nutlin-3a	-9.40 ± 0.42	-1.73 ± 0.51	-11.14 ± 0.12
Pip-2	-10.00 ± 0.10	0.90 ± 0.10	-9.10 ± 0.10
AMG-232	-10.30	-0.90	-11.20 ± 0.30
AM-7209	-10.90 ± 0.30	-1.35 ± 0.31	-12.25 ± 0.03
ITC measurements MDM2 long construct			
Nutfrag	-9.61 ± 0.98	2.79 ± 1.04	-6.81 ± 0.08
Nutlin-3a	-9.49 ± 0.27	-1.85 ± 0.38	-11.33 ± 0.16
Pip-2	-17.60 ± 0.20	6.80 ± 0.20	-10.80 ± 0.20
AMG-232	-17.68 ± 0.88	3.39 ± 0.47	-14.29 ± 0.41
AM-7209	-16.00 ± 0.22	0.44 ± 0.38	-15.54 ± 0.08

TABLE 5.2: Binding thermodynamic measurements for the MDM2 dataset. ΔG, ΔH and -TΔS are reported in kcal mol^{-1}. Standard deviations for MDM2-short/Nutfrag and MDM2-short/AMG-232 are not reported due to lack of replicates.

We also calculated the binding selectivity ($\Delta\Delta G$) between the two MDM2 constructs and the five inhibitors by deducting the binding free energy of the long construct from the free energy of binding of the lid absent region. The same approach was used to compute the differences in the change in entropy (T$\Delta\Delta S$) and in enthalpy ($\Delta\Delta H$). The above parameters are given in Table 5.3.

ITC measurements MDM2 short construct			
Ligands	$\Delta\Delta H$	$-T\Delta\Delta S$	$\Delta\Delta G$
Nutfrag	-4.60 ± 0.98	5.29 ± 1.04	0.69 ± 0.30
Nutlin-3a	-0.08 ± 0.50	-0.11 ± 0.63	-0.20 ± 0.40
Pip-2	-7.60 ± 0.22	5.90 ± 0.22	-1.70 ± 0.10
AMG-232	-7.38 ± 0.88	4.29 ± 0.47	-3.09 ± 0.30
AM-7209	-5.10 ± 0.37	1.79 ± 0.49	-3.28 ± 0.40

TABLE 5.3: Binding thermodynamic measurements for the MDM2 dataset. ΔΔG, ΔΔH and -TΔΔS are reported in kcal mol^{-1}.

Inspection of the results shows that the piperidinone class of compounds bind more favorably to the long construct of MDM2, especially in the case of AMG-232 (ΔΔG = -3.1 ± 0.3 kcal/mol) and AM-7209 (ΔΔG = -3.3 ± 0.4 kcal/mol). In addition, the enthalpic contributions to the binding free energy for Pip-2 and AMG-232 is almost identical between the full length and lid-truncated variant of MDM2 leading to the conclusion that the difference in the binding selectivity between these compounds is due to the entropic contributions. Indeed, the entropic cost required to order the flexible lid region of the apo structure is decreased in AMG-232 (-TΔΔS = 4.29 ± 0.47 kcal/mol) compared to Pip-2 (-TΔS = 5.90 ± 0.22 kcal/mol). On the contrary, AM-7209 has the lowest values of differences in the changes in entropy (TΔΔS = 1.79 ± 0.49 kcal/mol) and enthalpy (ΔΔH = -3.3 ± 0.4 kcal/mol).

For the nutlin inhibitors, ITC measurements show little (ΔΔG = 0.69 ± 0.30 kcal/mol for Nutfrag) or no difference (ΔΔG = -0.20 ± 0.40 kcal/-mol for Nutlin-3a) in binding between the two MDM2 constructs. This is in agreement with the experimental binding free energies measured by Michelsen *et al* for Nutlin-3a.[332] Interestingly, Nutfrag has a large unfavorable increase in entropy on binding to full lid MDM2 construct that mimics the thermodynamic signature of the piperidinones than Nutlin-3a. A possible reason for the absence of binding selectivity between short and long MDM2/nutlin complexes is that nutlins do not appear to structure the lid based on previous work of Michel group using computer simulations.[331]

We therefore sought next to compare the experimental findings with binding free energies calculated from the adaptive sampling protocol.

5.3.3 Adaptive sampling protocol results

A novel adaptive sampling protocol was employed for the computation of absolute binding free energies of the five MDM2/ligand complexes. We used an MDM2 variant where the disordered lid region was absent and two long constructs characterising different lid conformational states. We wanted to select the lid states that were more plausible for the given inhibitor based on the ITC assays. For this purpose, we used a lid state that the piperidinones order upon binding for the Amgen compounds, while for the nutlins we chose a more flexible extended closed lid state. All the convergence plots for the ten MDM2/ligand simulations together with the corresponding windows sampling plots are provided in the Appendix. The computed free energies of binding for each dataset and the standard state corrections are summarised in the following table:

absoluteFEP protocol MDM2 short construct		
Ligands	$\Delta G_{restraint}$	$\Delta G^{\circ}_{ModelC}$
Nutfrag	-4.89 ± 0.10	-6.24 ± 0.51
Nutlin-3a	-5.36 ± 0.40	-15.91 ± 0.91
Pip-2	-5.52 ± 0.16	-12.94 ± 0.59
AMG-232	-4.97 ± 0.20	-16.09 ± 0.86
AM-7209	-5.16 ± 0.28	-23.57 ± 0.75
absoluteFEP protocol MDM2 extended-closed construct		
Nutfrag	-4.56 ± 0.05	-7.47 ± 0.72
Nutlin-3a	-5.26 ± 0.22	-16.91 ± 0.33
Pip-2	-5.15 ± 0.13	-11.32 ± 0.79
AMG-232	-5.33 ± 0.12	-12.59 ± 1.22
AM-7209	-5.04 ± 0.16	-22.71 ± 1.41
absoluteFEP protocol MDM2 ordered construct		
Nutfrag	-4.45 ± 0.08	-8.95 ± 0.48
Nutlin-3a	-5.33 ± 0.19	-17.54 ± 1.07
Pip-2	-5.30 ± 0.15	-17.05 ± 0.70
AMG-232	-5.19 ± 0.20	-16.35 ± 0.58
AM-7209	-5.16 ± 0.21	-25.60 ± 0.91

TABLE 5.4: Results from all the absoluteFEP protocol simulations for the MDM2/ligand complexes. Energies are reported in kcal/mol

In addition, the computed and the experimental standard binding free energies of the short and the full-lid MDM2 variants are illustrated in Figure 5.12.

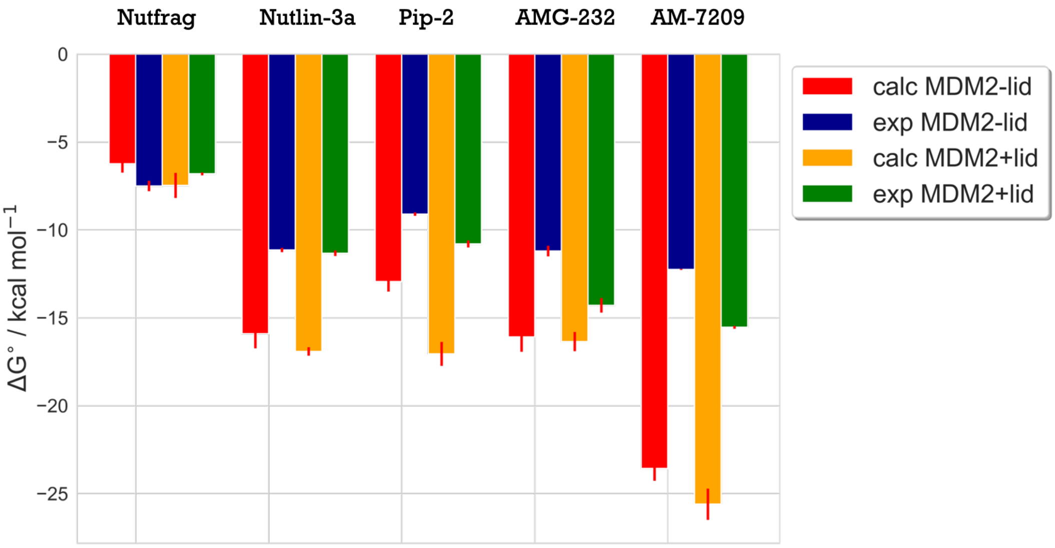


FIGURE 5.12: Computed and measured standard binding free energies for the MDM2 dataset (kcal/mol). The constructs used for the adaptive sampling protocol were the short MDM2 variant (MDM2-lid) and the preferred long MDM2 variant for each ligand (MDM2+lid).

The simulations seem to overestimate the binding free energy of all the ligands except Nutfrag. The greater the molecular weight, the greater the discrepancy. This could indicate a systematic issue with the force-field parameters describing the ligands. Another possible reason for this discrepancy is that the present calculations do not allow for extensive rearrangement of the protein. Since the calculations were all initiated from a well equilibrated complex this could artificially bias the results towards more negative free energies of binding. The effect could be expected to be greater for simulations of MDM2 constructs that include the flexible lid.

We also wanted to test the effect of the starting structure on the computed binding energies of each compound. For this purpose, we swapped the different lid conformations between the piperidinones and the nutlins. Then we repeated the absolute binding free energy calculations using the adaptive sampling protocol. All the convergence plots and windows sampling plots produced from these calculations are given in the Appendix. The standard binding free energies of each MDM2/ligand complex and the corresponding standard state correction terms are provided in the Table 5.4

Moreover, the predicted and the experimental standard binding free energies of the short and the full-lid MDM2 variants are shown in the Figure below:

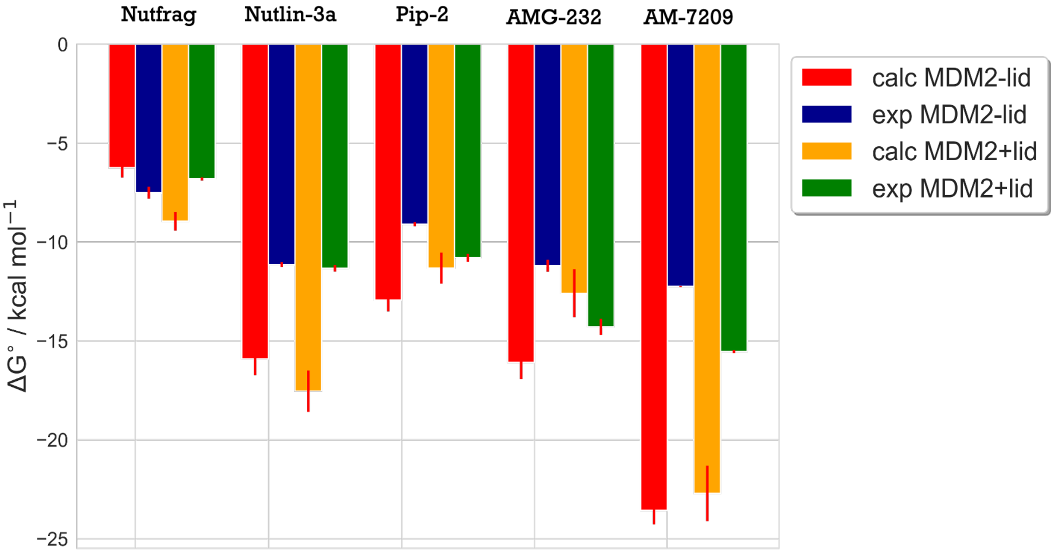


FIGURE 5.13: Calculated and measured standard binding free energies for the MDM2 dataset (kcal/mol). The protein conformations selected for the adaptive sampling protocol were the short MDM2 construct (MDM2-lid) and the swapped long MDM2 variant for each ligand (MDM2+lid).

The new standard binding free energies computed for the piperidinones were more positive compared to their preferred state. The binding affinity of AM-7209 is still overestimated compared to the experimental measurements, but Pip-2 and AMG-232 are in agreement with the ITC measurements. On the contrary, the effect of swapping protein conformations on the binding free energy of the nutlins was negligible. This effect was also examined by computing the trends in binding selectivity (difference in binding free energy between constructs that include/lack the lid region) for both conformations together with the difference in free energy of binding between the two lid states. The results for the whole dataset are summarised in Figure 5.14.

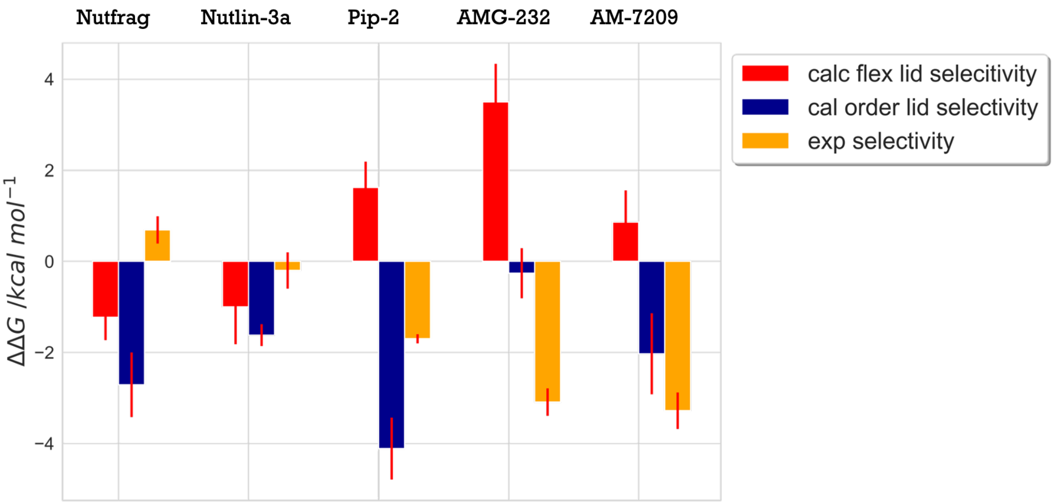


FIGURE 5.14: Difference in calculated free energy of binding between MDM2 constructs that include and lack the lid region. The binding selectivity for the flexible lid conformational state is coloured in red, while the binding selectivity for the ordered lid conformation is coloured in blue. Finally, the difference in measured binding free energy is shown in yellow color.

The trends in binding selectivity for the preferred lid state seem to be reproduced, with the possible exception of Nutfrag. Amgen compounds binds more strongly to the MDM2 construct with an ordered lid conformation than to the MDM2 short construct. They also bind less strongly to the MDM2 construct with a closed lid conformation than to the MDM2 short construct. This in agreement with Michelsen *et al* and Bueren-Calabuig *et al* which shows that the piperidones induce the formation of a α-helical/β-strand structure.([332],[331] Nutlin-3a does not show a significant preference for either lid conformations, which are slightly more preferred than the construct without lid. This is in line with Cesar's and Michelsen's ITC measurements that did not observe difference in binding free energy between long and short MDM2 constructs. Interestingly Nutfrag seems to show a slight preference for the ordered lid conformation. This could explain why the thermodynamic signature of Nutfrag shows a greater unfavorable entropic contribution on binding the MDM2+Lid construct than on binding the MDM2-Lid construct, whereas Nutlin-3a shows no such trend.

Another interesting observation is the convergence of the five different runs of each simulation as a function of time. From the convergence plots, provided in the Appendix, we can observe that only few systems such as MDM2-extended-closed-lid/Nutlin-3a had the five independent runs converged towards the average binding free energy at the end of the 50ns. The highest deviation between two independent runs was ca. 4 kcal mol^{-1} in the MDM2-extended-closed-lid/AM-7209 complex showing that unconverged free energies of binding can provide wrong conclusions for the binding selectivity of the systems. This proves the importance of running independent long simulations starting from different initial coordinates in order to get a better estimate of the free energy of binding of the system rather than relying on a single run of the protocol.

Overall the absolute binding free energy calculation protocol used here reveals differences in binding mechanisms between Nutlin and Piperidinone compounds that are in line with experimental trends.

5.3.4 Comparison between the adaptive sampling protocol and a docking protocol

We also sought to assess whether the experimental binding selectivity trends could be explained with simpler methods. For this purpose, we used a typical docking program provided by Flare and we calculated the docking scores of the 15 MDM2-ligand complexes. For consistency, we used the same input structures as for the adaptive sampling protocol. The computed binding free energies and the corresponding binding selectivity are provided in the Appendix. The binding selectivity plot of these plots are shown in Figure 5.15.

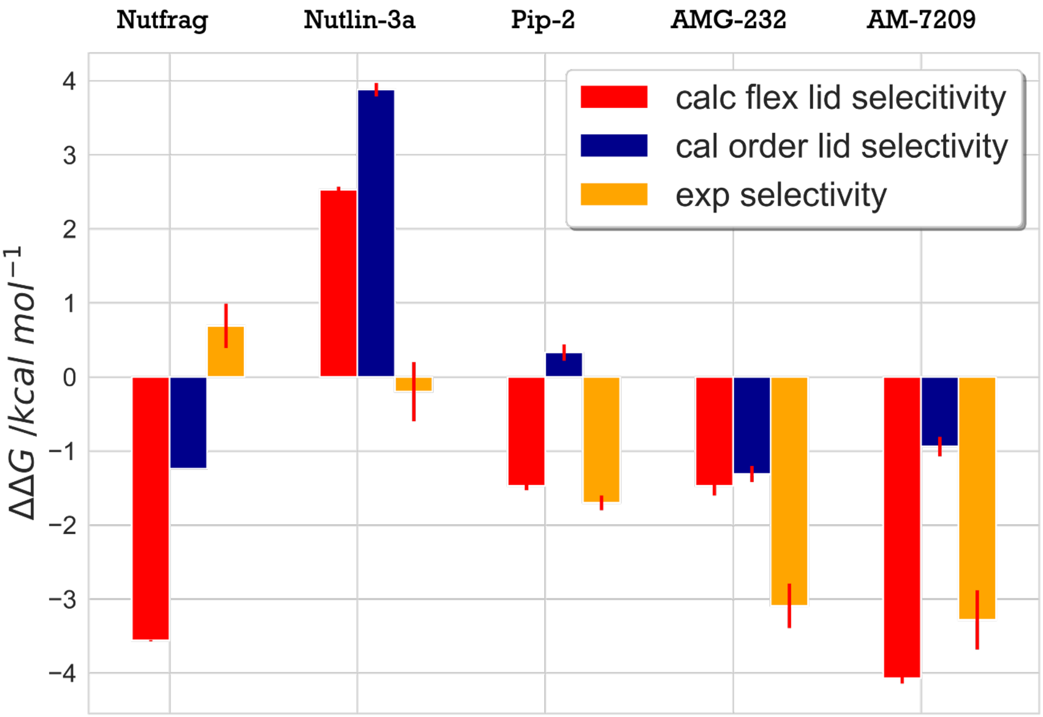


FIGURE 5.15: Difference in free energy of binding computed by docking studies between MDM2 variants that include and lack the lid region. The binding selectivity for the flexible lid conformational state is coloured in red, while the binding selectivity for the ordered lid conformation is coloured in blue. Finally, the difference in binding free energy measured by ITC assays is shown in yellow color.

Inspection of the results show that the experimental binding selectivity is not reproduced through the docking studies with the possible exception of Pip-2. Regarding to the piperidinone inhibitors, the binding selectivity computed from the flexible-lid MDM2 construct is closer to the experimental trends. Pip-2 and AMG-232 are predicted to have the same difference in free energy between variants that include/lack the lid region (-1.5 ± 0.1 kcal/mol and -1.5 ± 0.2 kcal/mol respectively), while AM-7209 shows a stronger binding selectivity (-4.1 ± 0.1) However, these results are in disagreement with Cesar's ITC assays that show similar binding preferences for the lid conformation between AM-7209 and AMG-232. In addition, the ordered-lid MDM2 variant is not predicted to be the preferred lid conformational state for the piperidinone class of compounds. Thus, the docking protocol is not able to correctly capture the tendency of Amgen compounds to structure the lid.

The docking studies are also unable to accurately predict the binding preference of the nutlin compounds. Nutfrag is found to bind preferably to the full lid MDM2 variants regardless the lid state (-1.24 ± 0.02 kcal/mol for structured lid and -3.56 ± 0.04 kcal/mol for the extended closed state), while for Nutlin-3a the binding selectivity is positive for the ordered (3.88 ± 0.03 kcal/mol) and the flexible lid state (2.5 ± 0.1 kcal/mol), indicating a strong preference for the short MDM2 construct. The docking results for the nutlin inhibitors are inconsistent with the ITC measurements that show no binding preference for the full-lid or the lid absent MDM2 constructs.

5.4 Conclusions

Alchemical free energy calculations were applied to estimate standard binding free energies of five lead-like inhibitors for the flexible protein MDM2. MDM2 contains a disordered lid region that adopts different conformations when bound to diverse ligands. A novel adaptive sampling protocol was established and validated to predict binding free energies of two full-lid MDM2 constructs and one lid absent MDM2 variant at a fraction of the computing cost of the original protocol developed in Chapter 3. The resulting free energies of binding were compared with calorimetry experiments performed by colleagues.

The adaptive sampling protocol was able to capture the tendency of structurally diverse ligands to bind a specific lid conformational state. The enhanced binding affinity of the piperidinone class of compounds for MDM2 constructs that include an ordered lid region is in accordance with the ITC experiments implemented by Cesar and Michelsen *et al* as well as previous work from Michel group using computer simulations.[332],[331] In addition, the protocol suggests that nutlin inhibitors show similar affinity for short or long MDM2 lid constructs with a small binding preference for extended closed lid conformations. This is also in agreement with the computational and experimental work performed from Bueren-Calabuig *et al.* and Cesar Mendoza Martinez respectively.

The absolute binding free energy protocol used here overestimates the free energy estimates of the entire dataset with few possible exceptions (MDM2-short/Nutfrag and MDM2-flexible-lid/AMG-232). A possible explanation for this discrepancy could be that the simulations ignore the 'lid reorganisation free energy' required to structure the lid in apo MDM2 from a disordered to a more ordered state. This energy is expected to be unfavorable and thus would more the binding free energy estimates more positive. This should be the focus of further work.

Finally, binding energies and binding selectivity profiles were computed by a typical docking program using Flare. The main finding is that this docking study was able to reproduce some binding selectivity trends (for instance Pip-2) but overall struggled to generate trends consistent with the binding preferences deduced from experimental observations.

Overall, the results demonstrate the important role played by the lid in controlling the potency of potency of p53/MDM2 inhibition by structurally diverse lead-like ligands, and outline a general strategy to target flexible protein regions with absolute binding free energy methods in structure-based drug design campaigns.

Chapter 6

Conclusions

Molecular Dynamics simulations (MD) and associated techniques could be beneficial for the hit-to-lead and lead optimization stage in the drug discovery process. Although MD-based methods have shown success within the pharmaceutical sector in the last couple of years, the sampling problem of these techniques is still a limitation for calculating the binding free energies of protein-ligand complexes. Therefore, this work has investigated sampling problem in the MD simulation methods by establishing protocols that compute the binding free energies of different biomolecular complexes.

Chapter 2 presented a relative free energy protocol using MD simulations as the sampling technique. This protocol was part of a novel computational workflow for the discovery of selective CypA and CypD inhibitors. The main target of this workflow was to modify compound 15 to extend deeper inside the *3 o'clock* pocket. For this purpose, a library of ca. 10000 analogues of compound 15 was constructed and all molecules were docked to CypD. The selection of desirable compounds for the next step of the protocol included filtering of these molecules according to docking scores, synthetic feasibility and structural diversity. Then MD simulations were employed to identify molecules that could maintain stable interactions within the *3 o'clock* pocket. Finally, relative FEP calculations were able to suggest auspicious designs that could improve potency and selectivity for the two Cyclophilin isoforms. Disulfide derivatives offer substantial binding and selectivity for CypA over CypD, whereas MP030 and the tetrahydropyran analogue were predicted to bind more strongly to CypD over CypA. In addition, MD/FEP methods led to the discovery of piperidine and morpholine scaffolds that could replace the bromine of compound 15 and improve its physicochemical properties as well as bicyclic and pyrimidine analogues that could reduce its toxicity.[177] Therefore, the workflow was able to inform the synthesis of second generation tri-vector inhibitors, that will be characterised using biophysical techniques.

Chapter 3 addressed a more difficult task: the generation of an absolute binding free energy protocol, using MD simulations as a sampling technique, for the prediction of absolute free energies of binding of protein-ligand complexes. The efficiency of this protocol was evaluated in the SAMPL6 challenge. In the context of this challenge, series of blinded predictions of standard binding free energies were made with the SOMD software for a dataset of 27 host–guest systems featuring two octa-acids hosts (*OA* and *TEMOA*) and a cucurbituril ring (*CB8*) host. For this purpose, three different models were used, *Model A* calculated the free energy of binding based on a double annihilation protocol; *Model B* added the long-range dispersion and standard state corrections and *Model C* introduced an empirical correction term derived from a regression analysis of SAMPL5 predictions previously made with SOMD. The performance of each model was evaluated with two different setups; *buffer* explicitly matched the ionic strength from the binding assays, whereas *no-buffer* neutralized the host–guest net charge with counter-ions. The results obtained from our protocol were ranked among the top ranked submissions in terms of accuracy and correlation with experimental data. *Model C/no-buffer* showed the lowest MUE for the overall dataset (MUE 1.29 < 1.39 < 1.50 kcal mol^1, 95% CI), while the *buffer* setup improved significantly results for the CB8 host only. Correlation with experimental data ranges from poor for *CB8* (R^2 0.04<0.12<0.23), to excellent for the host *TEMOA* (R^2 0.91<0.94<0.96). Finally, the SAMPLing challenge examined the level of agreement between seven free energy methods applied to three host-guest molecules, namely OA-G3, OA-G6 and CB8-G3, that were parameterised with the same force-field. The analysis framework devised by the organisers showed compelling differences in the converged absolute binding free energies for the different methods ranging from 0.3 to 1 kcal mol^{-1}. For instance, comparison of the YANK's energies provided by the organizers with our calculations demonstrated a negligible difference (OA-G6 0.2 0.2 kcal mol^{-1}), a modest difference (OA-G3, 1 0.2 kcal mol^{-1}), and a significant difference (CB8-G3, 3 0.7 kcal mol^{-1}). Taking everything into consideration, there is hope that the AFE protocol could be further enhanced for more difficult protein-ligand complexes.

Chapter 4 presented two MD simulation protocols for the computation of the standard binding free energy of the ligand 10058-F4 to the intrinsically disordered protein c-Myc, and the characterisation of the binding mechanism of this molecule with c-Myc. The first protocol used the AFE protocol developed in chapter 3 and gave reproducible results but unfortunately the calculated standard binding free energy (ca. -2 kcal.mol^{-1}) diverged significantly from the van't Hoff analysis of Heller *et al* (ca. -6 kcal.mol^{-1}).[289] The second protocol combined extensive MD simulations and Markov State models in order to model the binding process of 10058-F4 to c-Myc. The MSM models were able to define binding as a two-state process using as features distances between parts of the ligand and the Ca's of c-Myc. In addition, they highlighted the N-terminus residues of the oncoprotein as critical for the binding process and the interactions were mainly hydrophobic. This result was in agreement with a previous study from Cuchillo and Michel, which used the BEMD method to sample the energetic landscape of c-$\text{Myc}_{402\breve{}412}$/10058-F4 complex.[290] However, the binding free energies obtained from the MSM models were unable to reproduce the experimental values, and were broadly similar to those obtained with the AFE protocol. Finally, the oncoprotein c-Myc was parameterised with three different force-fields (AmberIDP, Charmm36m, FF14SB) to assess to what extent the MSM results depend on the chosen forcefield. None of the three force-fields showed significantly different standard binding free energy. .

Finally, chapter 5 introduced an adaptive sampling version of the AFE protocol to estimate standard binding free energies of five lead-like inhibitors for the flexible protein MDM2. MDM2 also contains an intrinsically disordered region that can adopt multiple conformations upon binding to structurally-diverse inhibitors. The novel adaptive sampling protocol was employed to compute the absolute binding free energies of two full-lid MDM2 variants and one lid absent MDM2 construct using approximately one fifth of the computational time of the original protocol. The protocol was also able to predict the binding preference of the piperidinones for the ordered lid conformation as well as the similar affinity of nutlins for short or long constructs with a small preference for the more disordered lid region. These results were in agreement with ITC experiments carried out by lab member Cesar Mendoza-Martinez, and published work from Michelsen *et al*, as well as with the computational work performed previously by Bueren-Calabuig *et al*.[332],[331] However, the binding free energies computed from the adaptive sampling protocol were generally overestimated for all the MDM2/ligand complexes. This could be happening because in the alchemical free energy calculations the lid moves very little, whereas we know that is is quite flexible in practice. Thus, further work can be performed to devise a correction term ('lid reorganisation free energy') that would make the binding free energy estimates more positive. Taking everything into account, the results were encouraging as the protocol was able to predict the binding selectivity of the different MDM2 inhibitors at a fraction of the computing cost of the original AFE protocol.

Overall, this work has pushed the boundaries of binding free energy calculation methods for flexible protein-ligand complexes, and it is hoped that some of the methodologies tested during this research will find broader applications in pharmaceutical RD in due course.

Appendix A

Introduction

A.1 Total momentum and energy conservation.

The mathematical prove of the total momentum and energy conservation is described below:

1. Total momentum, P, which is the sum of all the momenta of N number of particles, must be fixed, therefore it must not depend on time, as it is illustrated in Equation A.1:

$$\frac{dP}{dt} = \sum_{i=1}^{N} \frac{dp_i}{dt} = \sum_{i=1}^{N} f_i = \sum_{i=1}^{N} \sum_{j \neq 1}^{N} f_{ij} = f_{12} + f_{21} + f_{13} + f_{31} + ... = 0 \tag{A.1}$$

   where the total force of a particle i is calculated as the sum of all the forces due to all the other particles, $\sum_{i=1}^{N} \sum_{j \neq 1}^{N} f_{ij}$. Newton's third law of motion states that every action has a reaction, f$_{ji}$ = f$_{ij}$ thus, as the forces are vectors and operate in opposite directions, the total sum of the forces of all the particles is equal to zero.

2. Total energy, E, must be also constant and must not depend on time, as it is depicted in Equation A.2:

$$\begin{aligned}\frac{dE}{dt} = \frac{d}{dt}(K+U) = \frac{d}{dt}(\sum_{i=1}^{N}\frac{p_i^2}{2m_i} + \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq 1}^{N}u_{ij}) = \\ = \sum_{i=1}^{N}(\frac{2p_i}{2m_i}\frac{dp_i}{dt} + \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq 1}^{N}(\frac{dq_i}{dt}\frac{du_{ij}}{dq_i} + \frac{dq_j}{dt}\frac{du_{ij}}{dq_j}) = \\ = \sum_{i=1}^{N}v_i f_i - \frac{1}{2}\sum_{i=1}^{N}v_i f_i - \frac{1}{2}\sum_{i=1}^{N}v_i f_i = 0 \quad \text{(A.2)}\end{aligned}$$

Where K is the kinetic energy of the particles and U is the potential energy of the particles.

A.2 Ewald summation method

The Ewald method for the computation of the long-range interactions is presented below: Each ion is considered to interact with all the other replicas of all the other ions besides itself. This is a computationally costly simulation, because there is a need to sum all the Coulombic interactions between all the replicas of the simulation cell. For a system with N particles that are assumed to be located in a cube with edge length L and a conducting boundary $\epsilon = \infty$, Coulombic interactions are expressed by Equation A.3:

$$U(Coulomb) = \frac{1}{2}\sum_{n}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{|r_{ij} + L_n|} \quad \text{(A.3)}$$

where n are the lattice vectors, $n = (n_x L, n_y L, n_z L)$ and $n_x, n_y, n_z = 0, \pm 1, \pm 2 \pm \ldots$ q_i and q_j are the charges of the ions and r_{ij} is the distance between the ions. The factor $\frac{1}{2}$ avoids the double counting. That is a very large sum and it is difficult to converge in a reasonable computation time. However, there is a way to split up that sum into two quickly converging contributions by using properties of Fourier transforms in Equation A.4:

$$U(Coulomb) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j erf_c(\sqrt{\alpha} r_{ij}}{r_{ij}} + \frac{1}{2\pi L^3}\sum_{k\neq 0}\frac{4\pi}{k^2}e^{-frack^2 4\alpha^2} p_q(k)p_q(-k) - \\ -\frac{\alpha}{\sqrt{\pi}}\sum_{i=1}^{N} q^2 \tag{A.4}$$

Where the first term is the real space sum that involves the complementary error function from the statistics. The second term is the sum at the reciprocipal space, which is due to a charge distribution $p_q(k)$ that consists of a periodic sum of Gaussians where k is the wave vector and is equal to $\frac{2\pi n}{L^2}$. In addition, α is a screening parameter that equals to $\frac{5}{L}$.

Appendix B

Computationally Driven Discovery of Novel Cyclophilin A and D Inhibitors

B.1 Binding mode I and II

Results from the extensive MD simulations performed by de Simone et al on CypA in complex with compound **1**.[156]

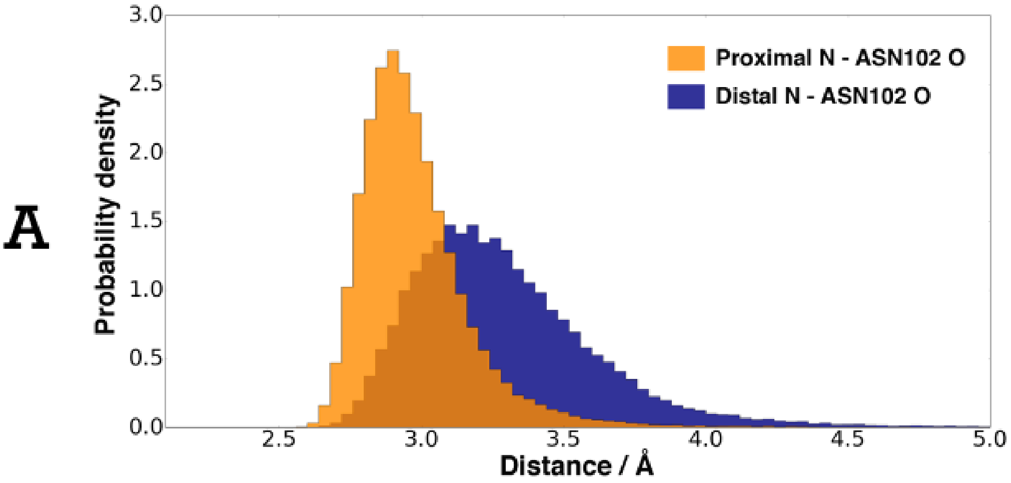


FIGURE B.1: Distance distribution between the two NH in the urea. Orange is used for the proximal one and blue for the distal one. Adapted from [156]

Schematic representation of the binding II mode hypothesis conducted by the Michel group.[156]

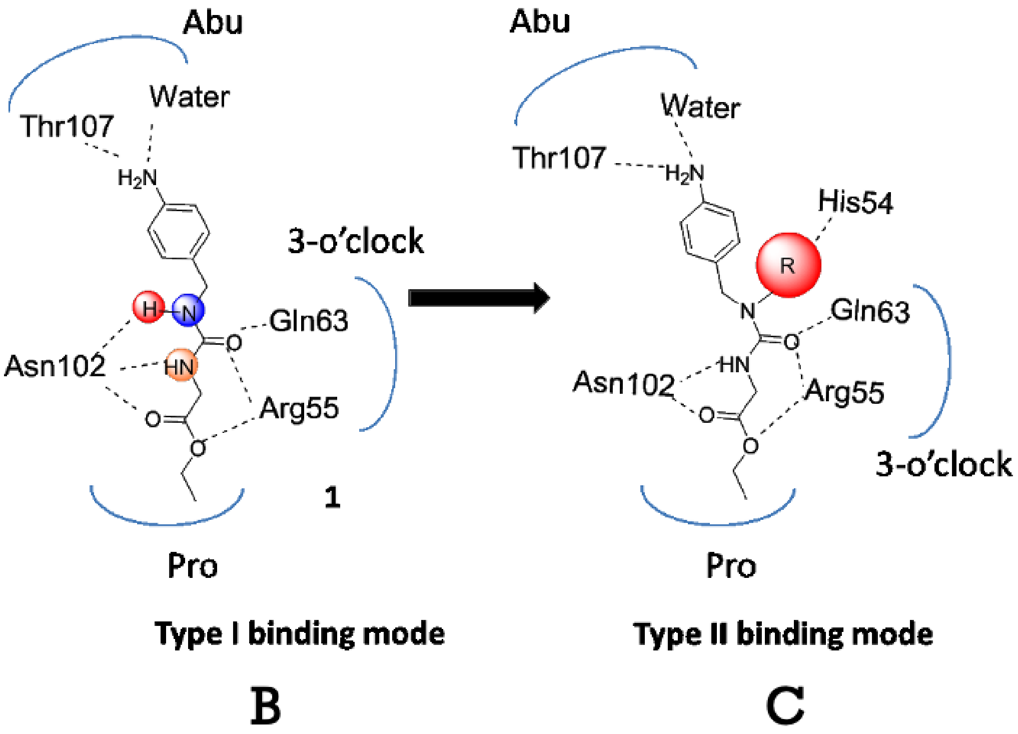
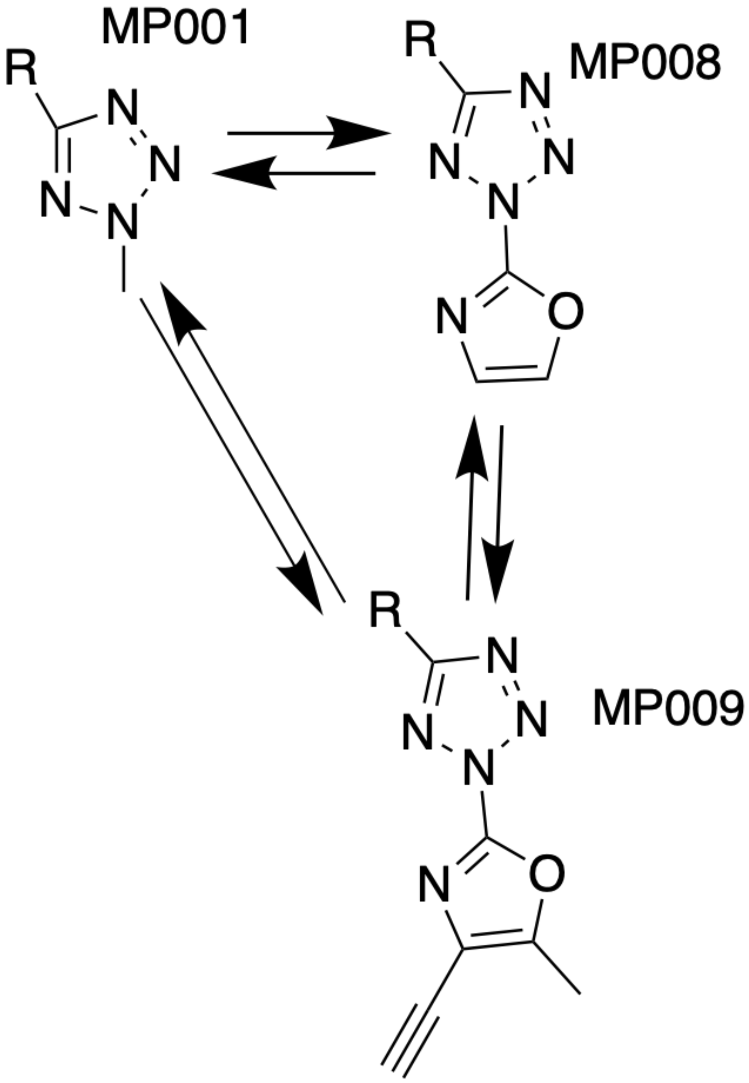
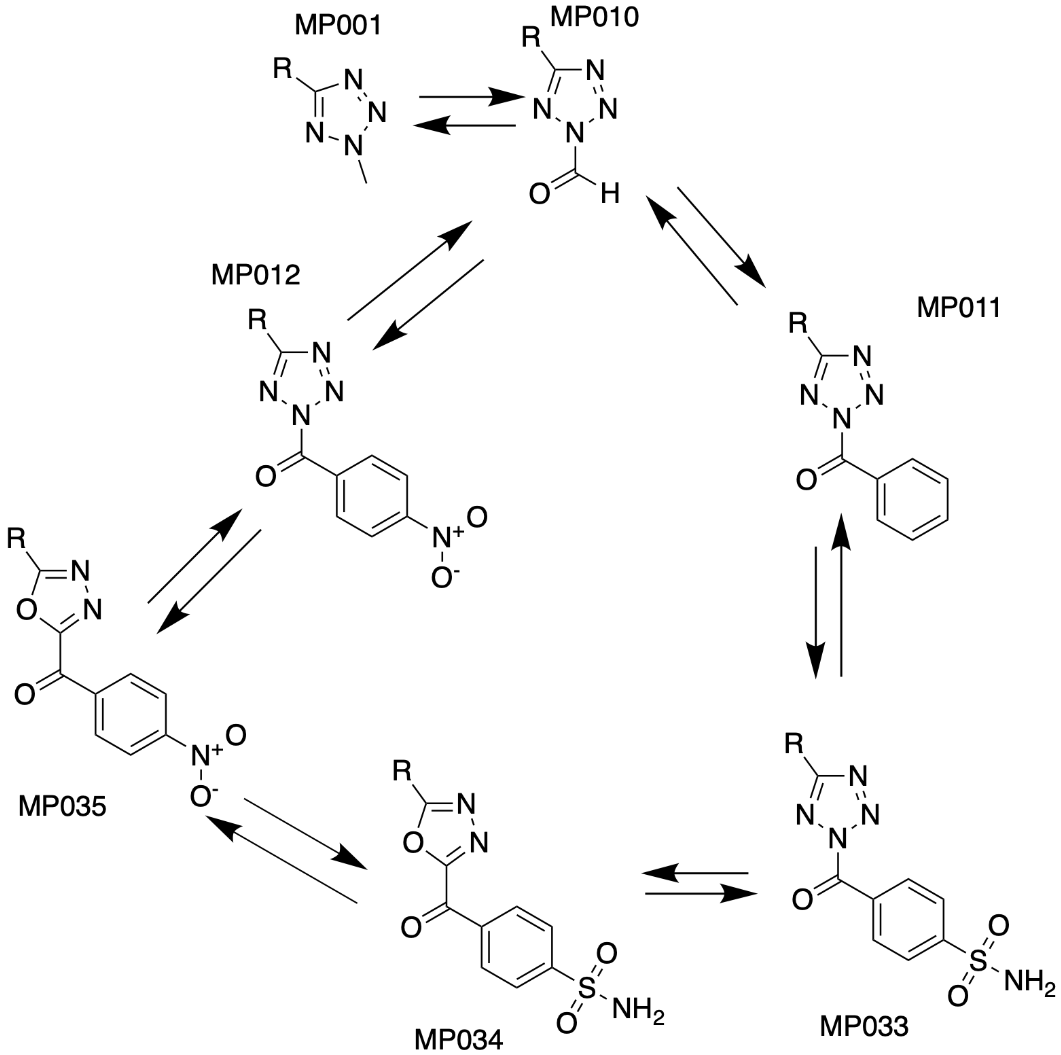
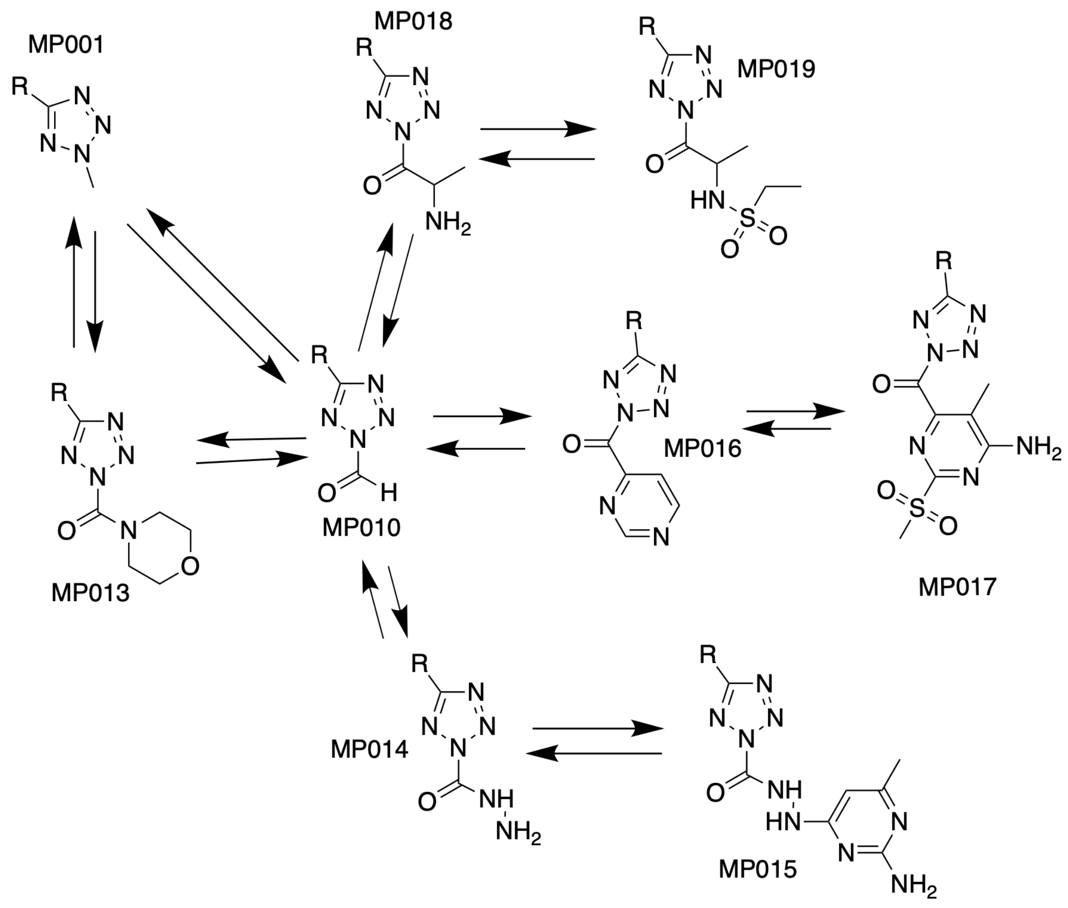


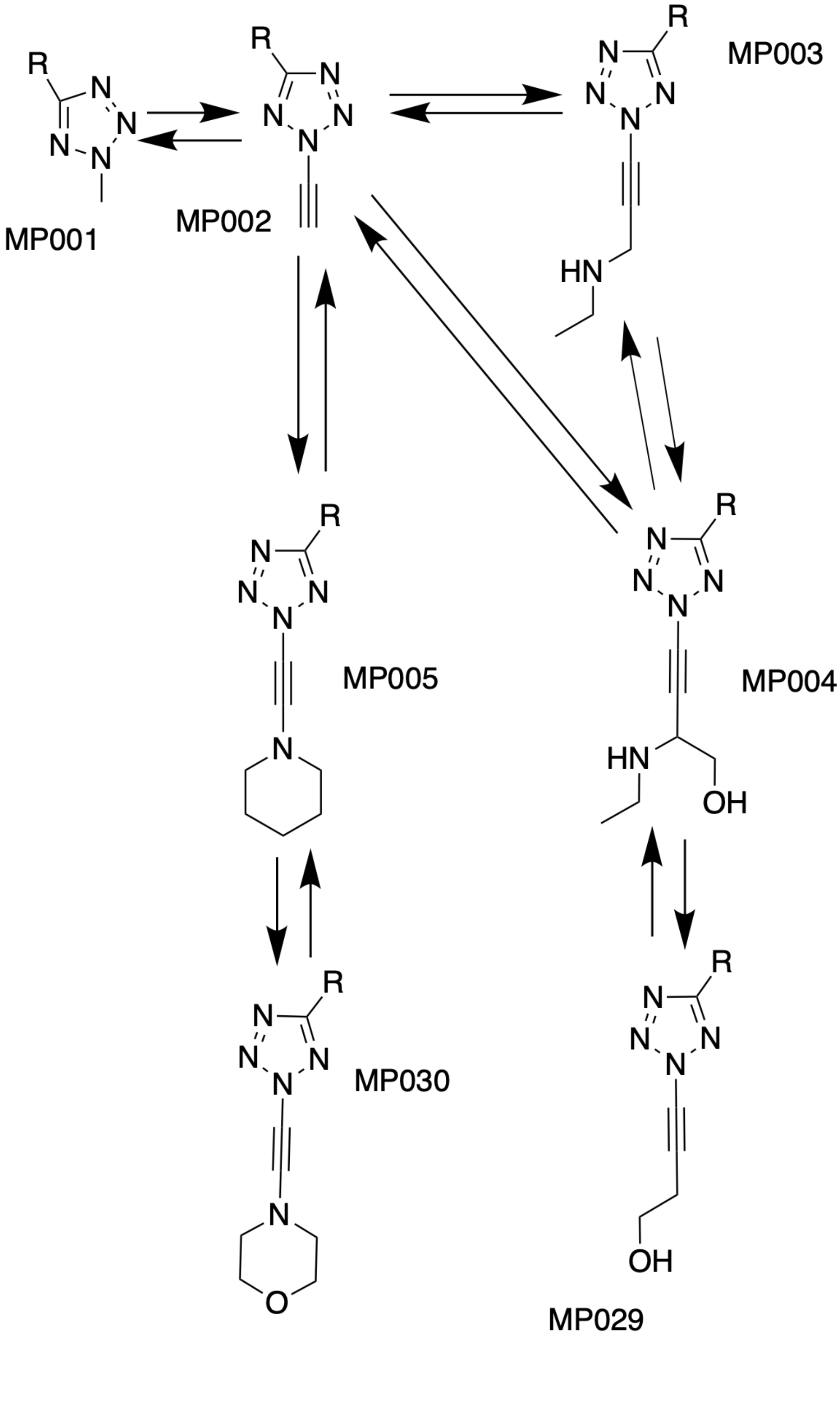
FIGURE B.2: **B)** Interactions between compound **1** and CypA in a type-I binding mode. **C)** Interactions between a hypothetic acylated urea and CypA in a type-II binding mode. Adapted from [156]

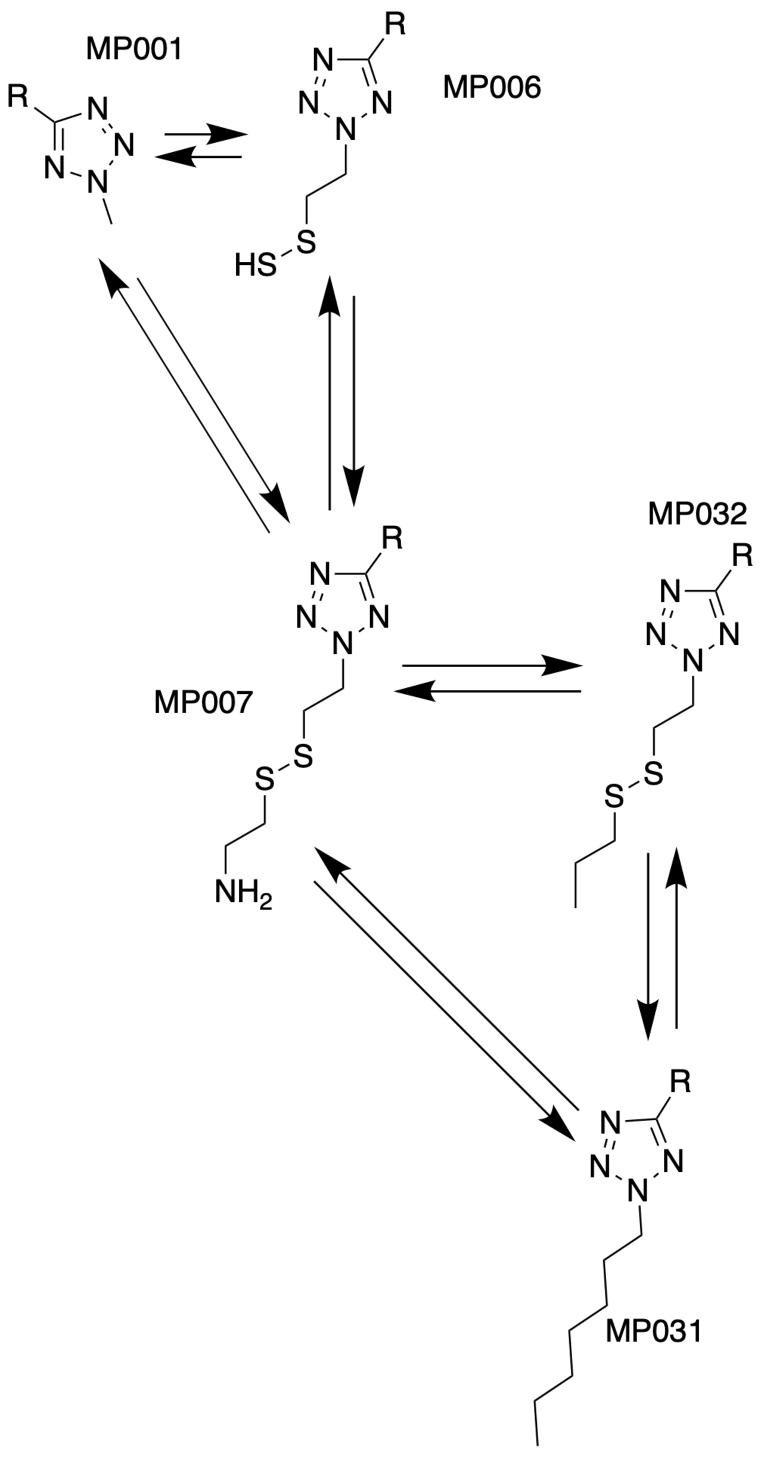
B.2 Perturbation maps for free energy calculations

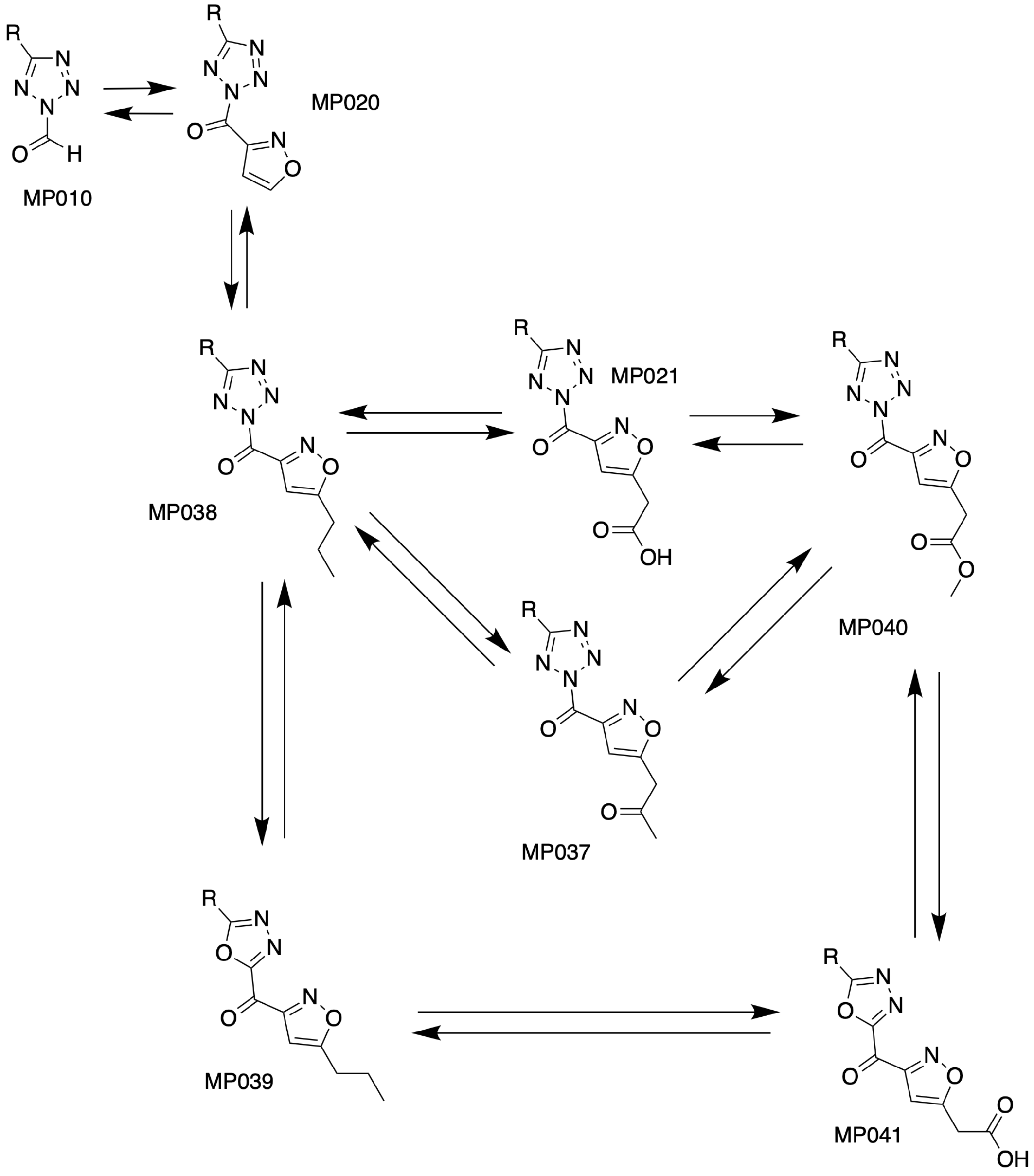


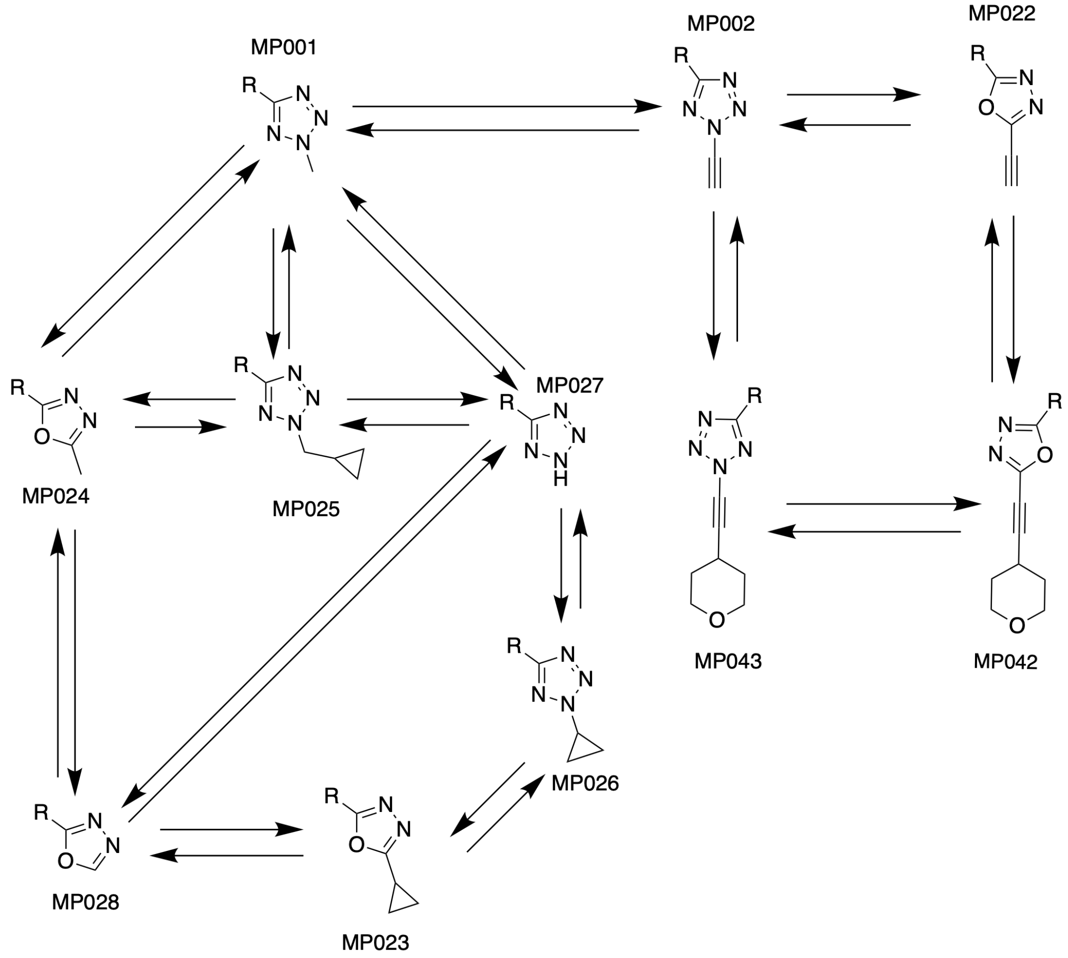

MP001
MP010
MP012
MP011
MP035
MP034
MP033

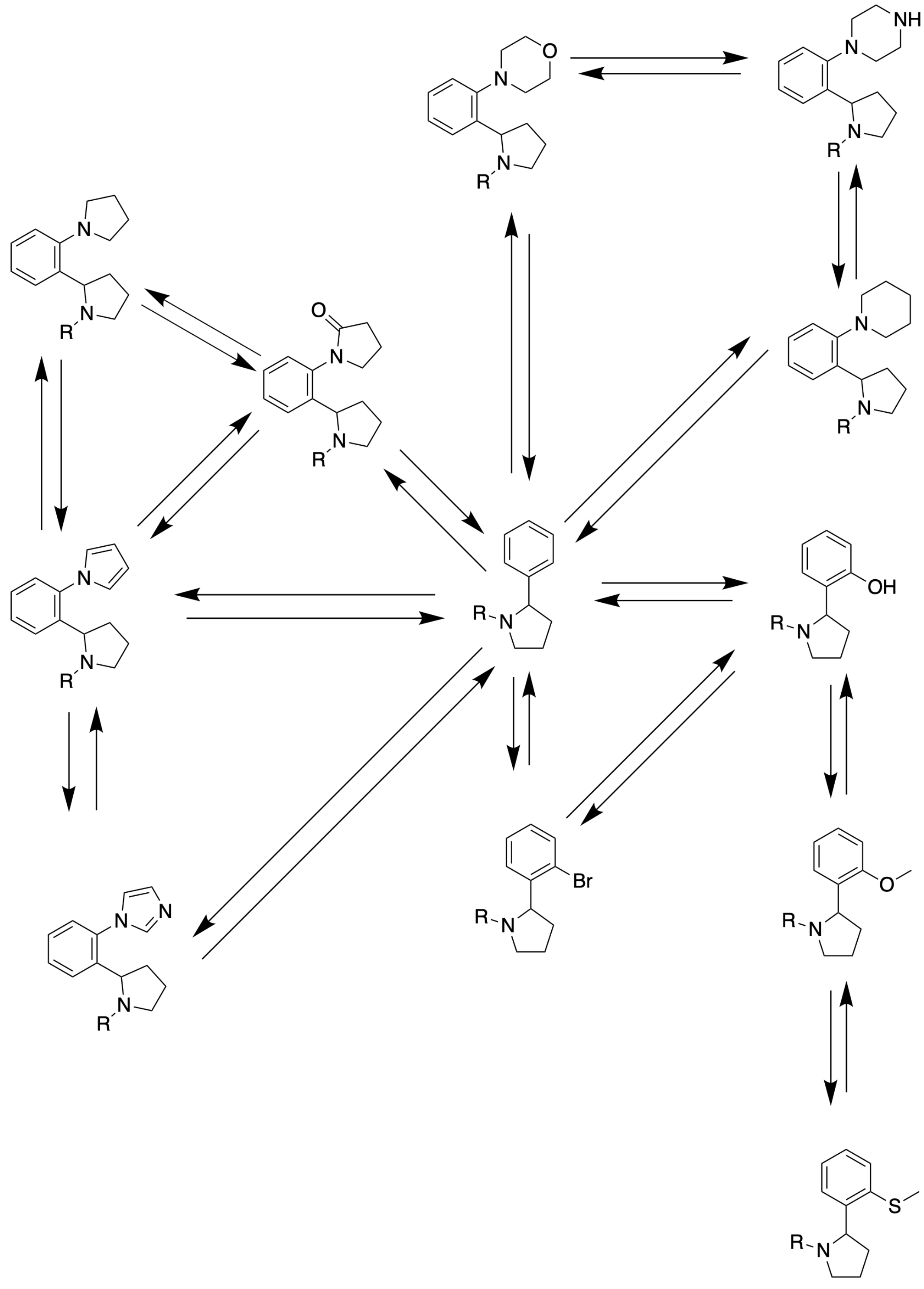

MP001
MP018
MP019
MP013
MP010
MP016
MP017
MP014
MP015

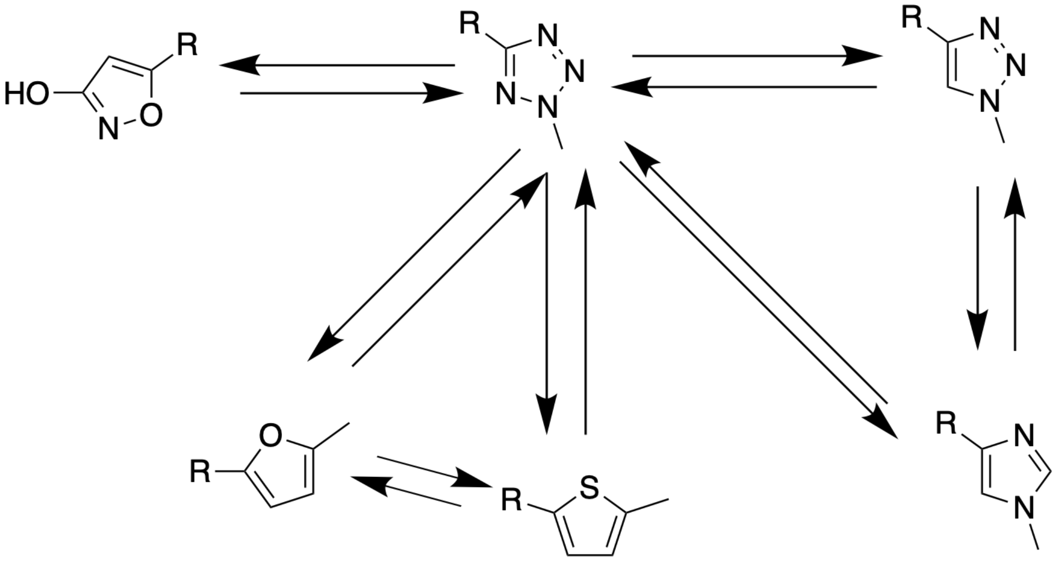

MP001
MP002
MP003
MP004
MP005
MP029
MP030


MP001
MP006
MP007
MP032
MP031


MP010
MP020
MP038
MP021
MP040
MP037
MP039
MP041


MP001
MP002
MP022
MP024
MP025
MP027
MP043
MP042
MP026
MP028
MP023


O
NH
N
R
OH
Br
S


R
HO
N
O
R
N
N
N
N
R
N
N
N
R
O
R
S
R
N
N

Appendix C

Blinded Predictions of Standard Binding Free Energies: Lessons Learned from the SAMPL6 Challenge

C.1 α, β, R^2 parameters for the *Model C* of SAMPL6 challenge

Host	α	β	R^2
OA	+1.32	-0.27	0.87
TEMOA	+1.10	-1.29	0.77
CB8	+1.96	+1.70	0.64

TABLE C.1: α, β, R^2 parameters for the *Model C* of SAMPL6 challenge. The parameters were calculated from the linear regression models performed by correlating the SAMPL5 binding free energies calculated with SOMD to experimental data. α and β are the slope and intercept of the linear regression model and R^2 is the coefficient of determination of the model.

Appendix D

Prediction of Absolute Binding Free Energies of Ligands for the Intrinsically Disordered Protein c-Myc

D.1 ITS plots

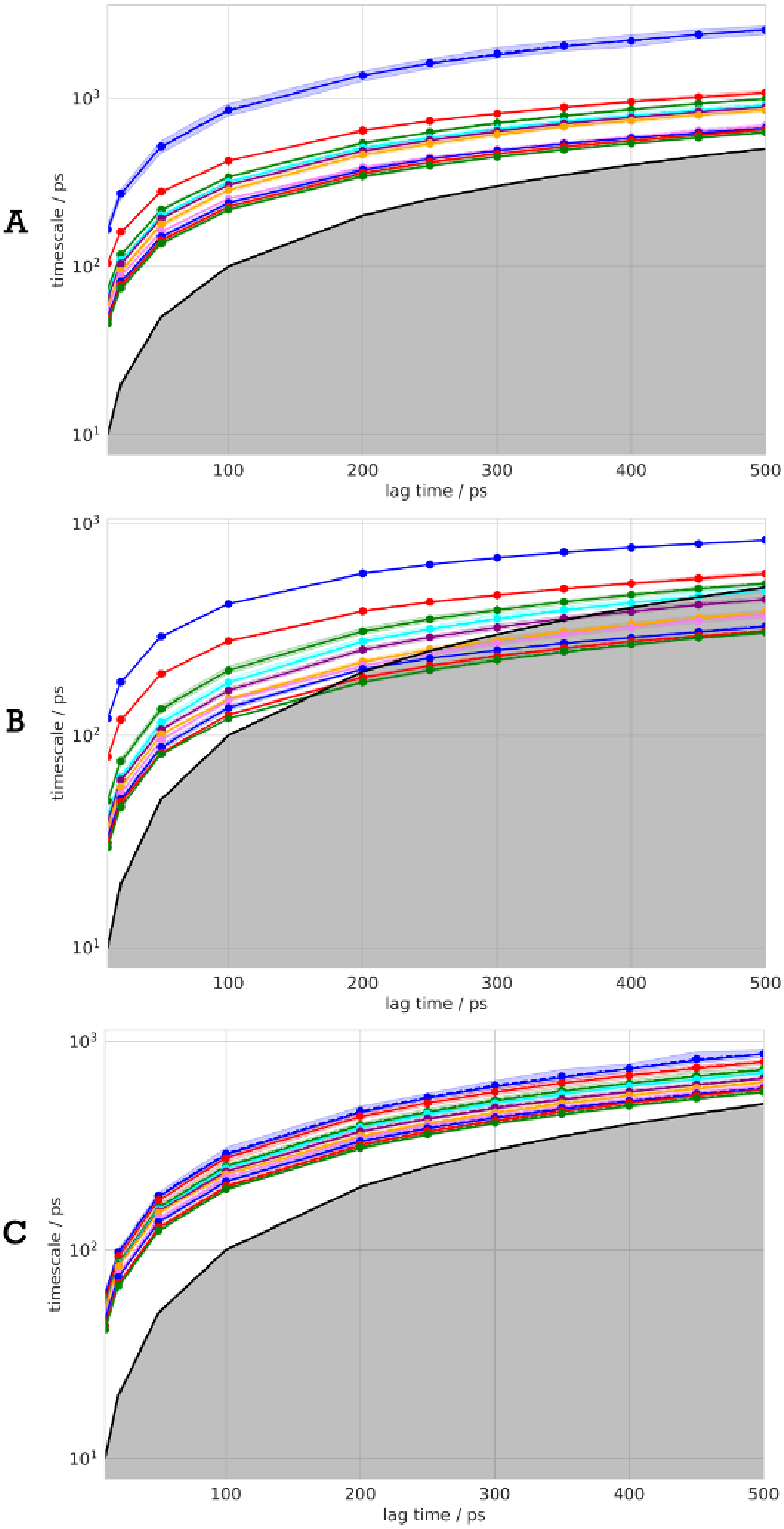


FIGURE D.1: Implied time scales plots of the first metric for the three force fields, **A** AmberIDP, **B** Charmm36m, **C** FF14SB.

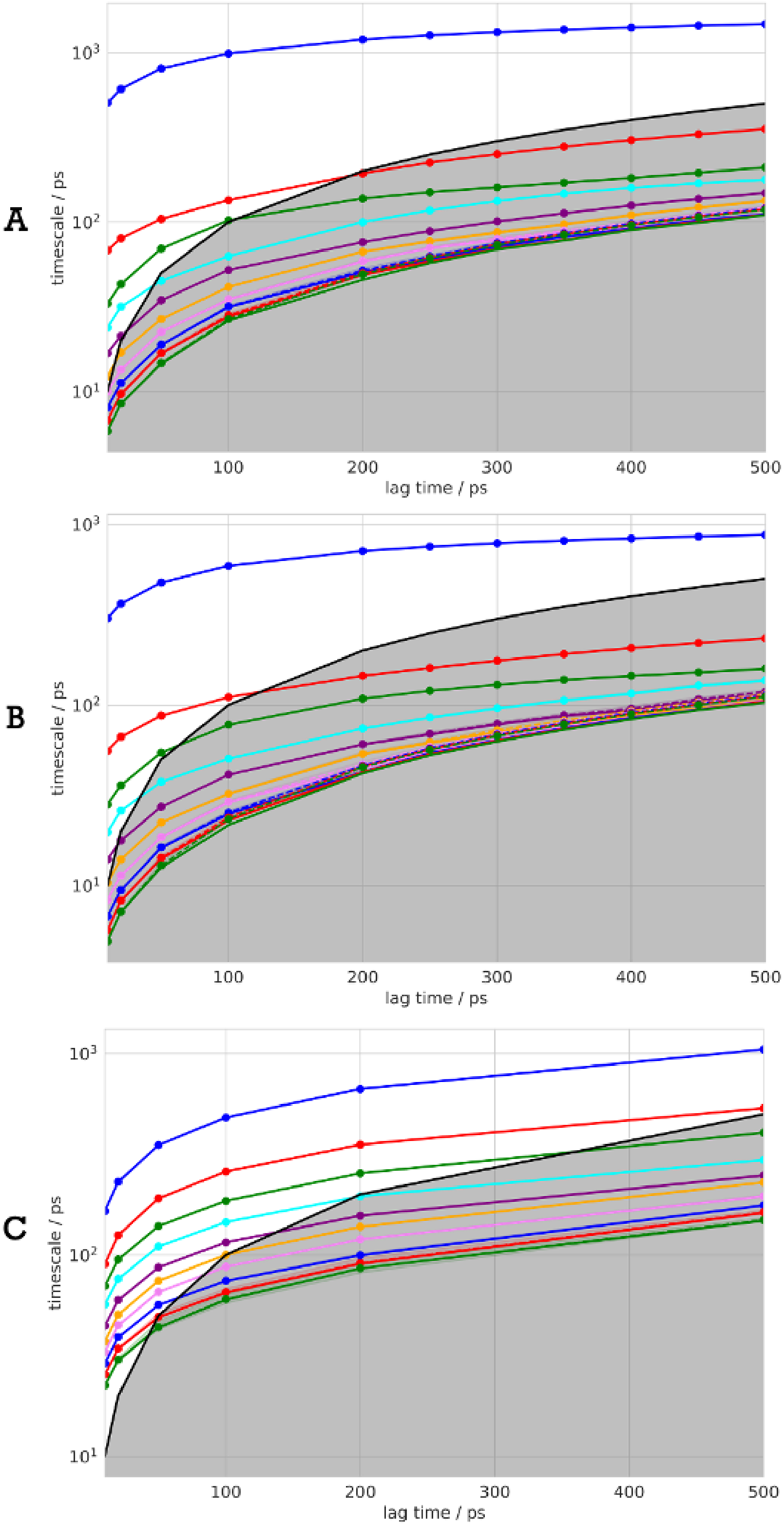


FIGURE D.2: Implied time scales plots of the third metric for the three force fields, **A** AmberIDP, **B** Charmm36m, **C** FF14SB.

D.2 CK tests

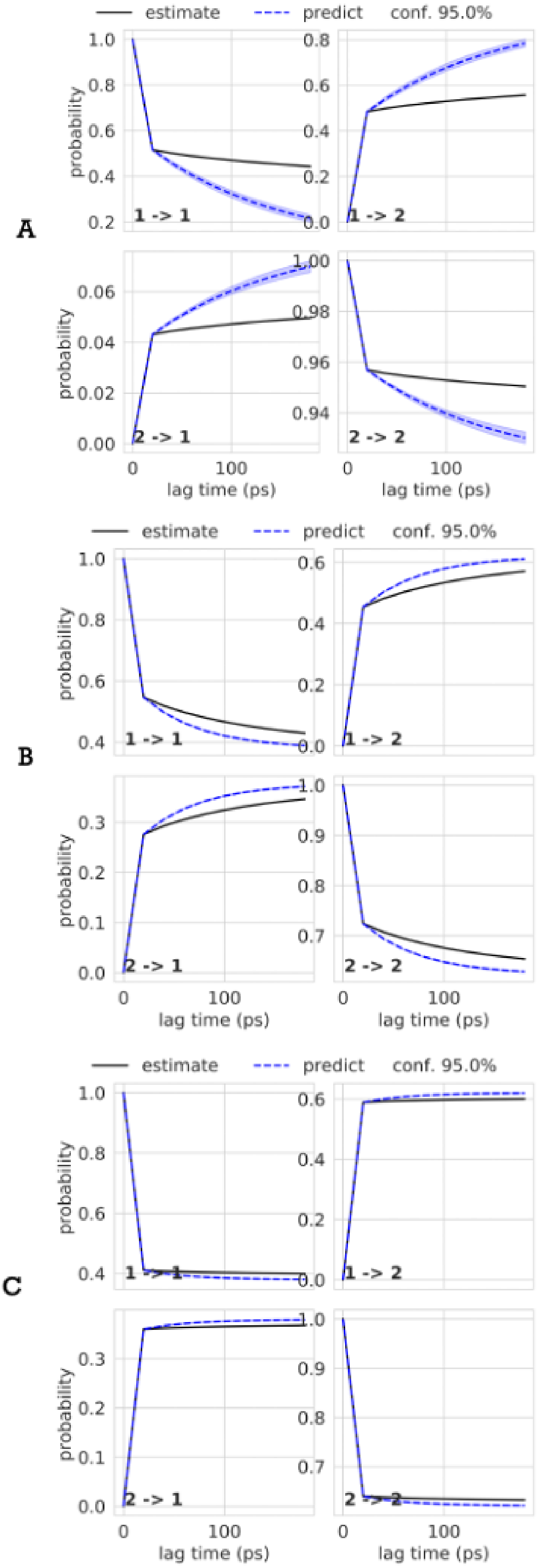


FIGURE D.3: Chapman-Kolmogorov test plots of the first metric used for the three force fields, **A** AmberIDP, **B** Charmm36m, **C** FF14SB.

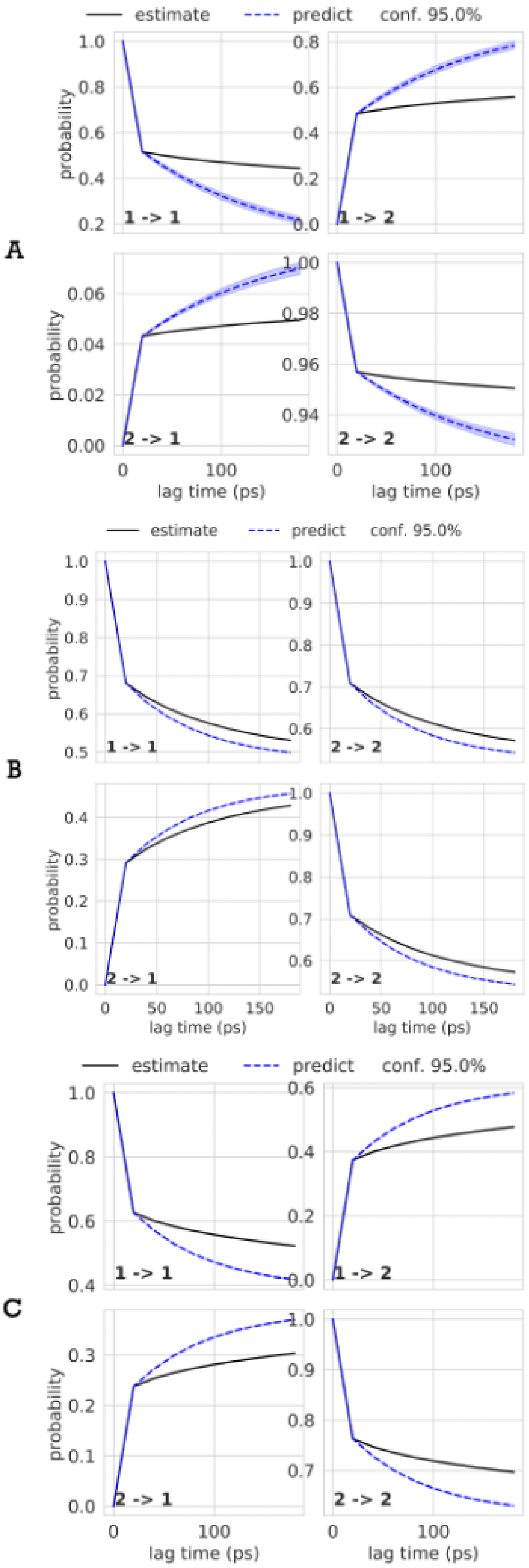


FIGURE D.4: Chapman-Kolmogorov test plots of the third metric used for the three force fields, **A** AmberIDP, **B** Charmm36m, **C** FF14SB.

D.3 Probability distribution of distances

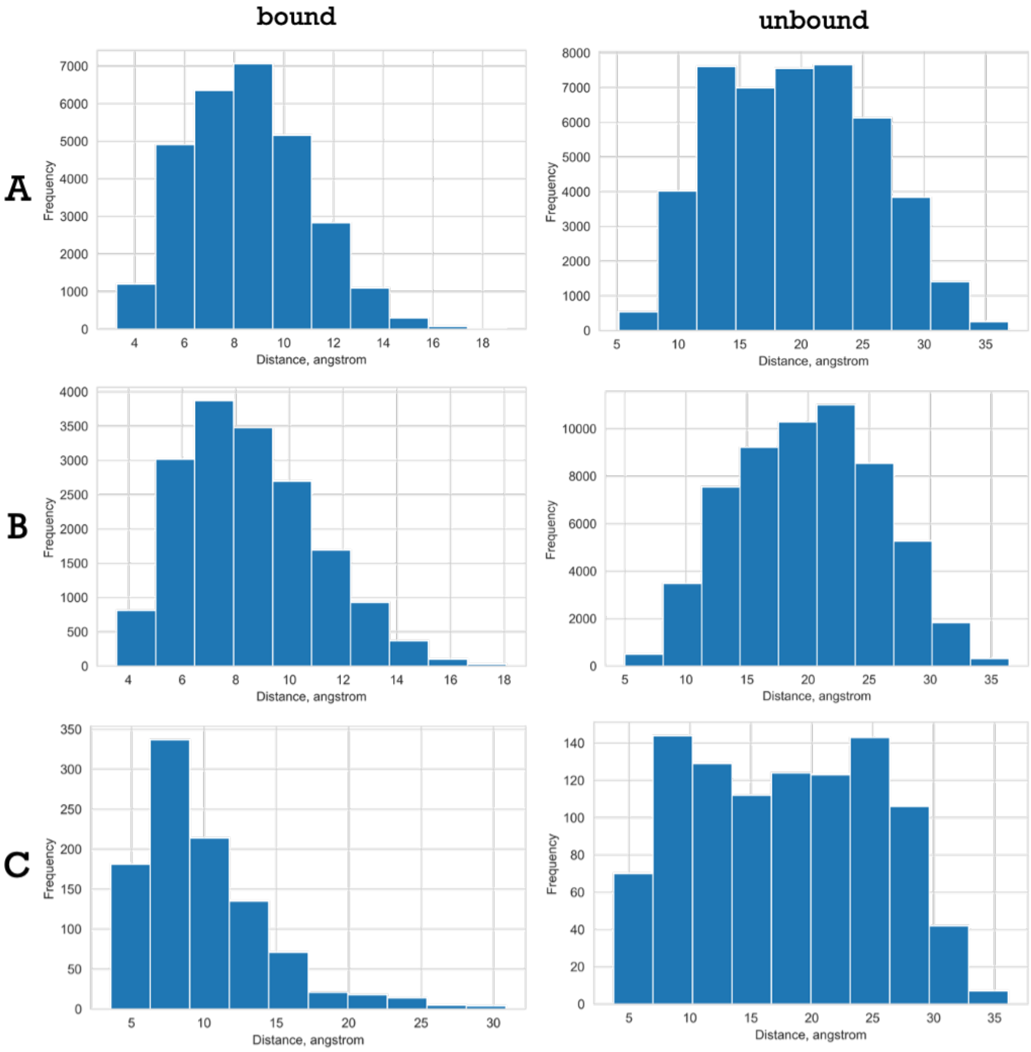


FIGURE D.5: Probability distribution of distances of the bound and unbound states of the first metric applied for the three force fields, **A** AmberIDP, **B** Charmm36m, **C** FF14SB.

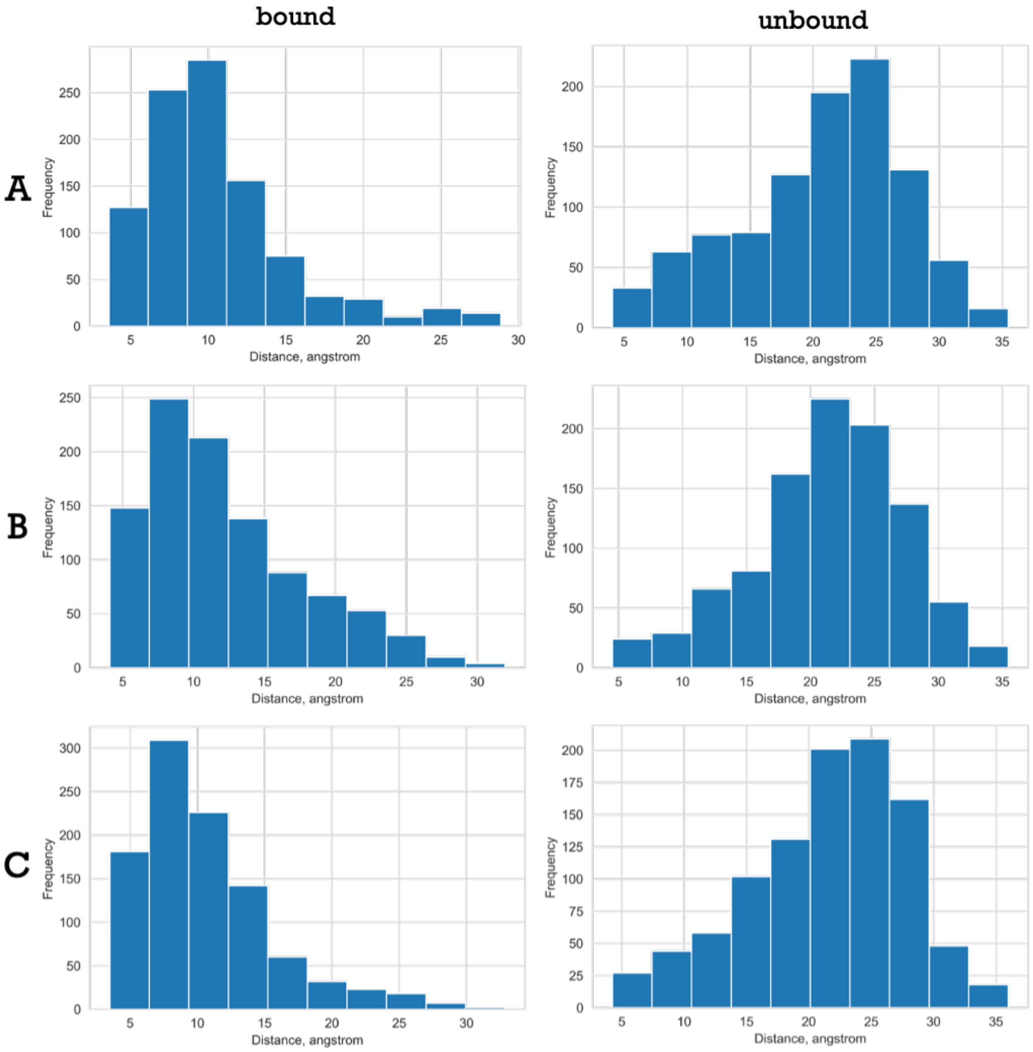


FIGURE D.6: Probability distribution of distances of the bound and unbound states of the third metric applied for the three force fields, **A** AmberIDP, **B** Charmm36m, **C** FF14SB.

D.4 Total number of contacts for 10058-F4

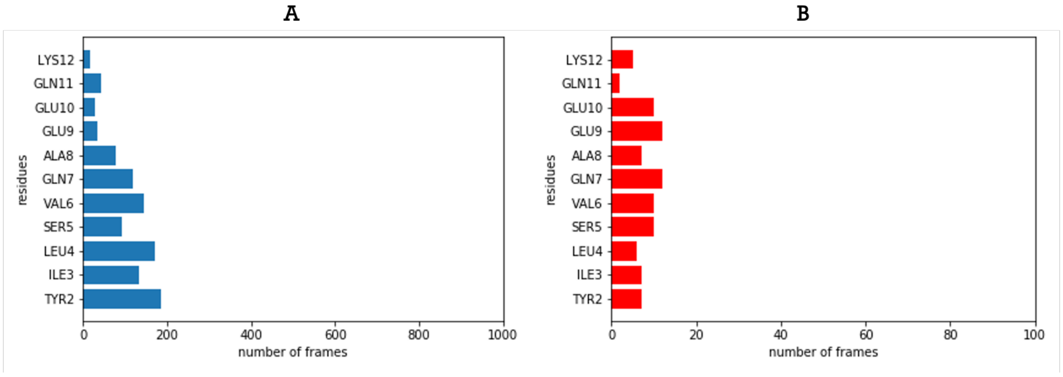


FIGURE D.7: A) Total number of hydrophobic contacts (on the left with blue color) and B) Number of hydrogen bonds (on the right with red color) for Charmm force-field. The x-axis has been truncated to 100 for visualisation purposes (number of contacts less than 100.

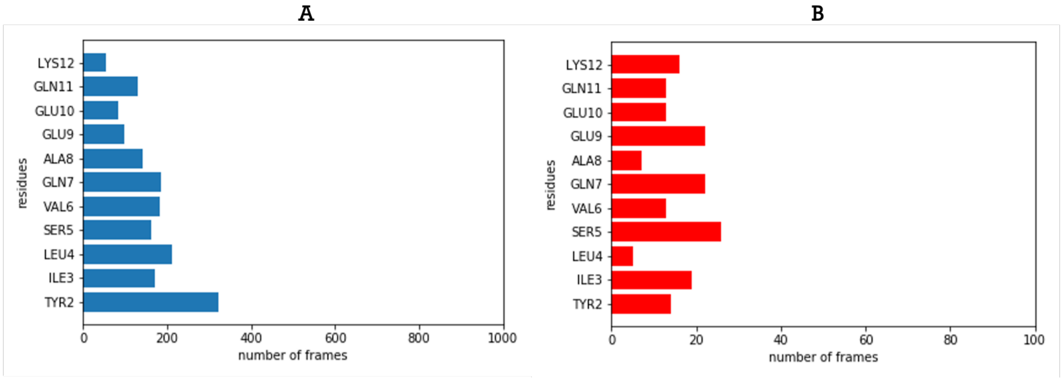


FIGURE D.8: A) Total number of hydrophobic contacts (on the left with blue color) and B) Number of hydrogen bonds (on the right with red color) for ff14SB force-field. The x-axis has been truncated to 100 for visualisation purposes (number of contacts less than 100.

Appendix E

Absolute Binding Free Energy Calculations of Ligands for the Flexible Protein MDM2

E.1 ITC titration data

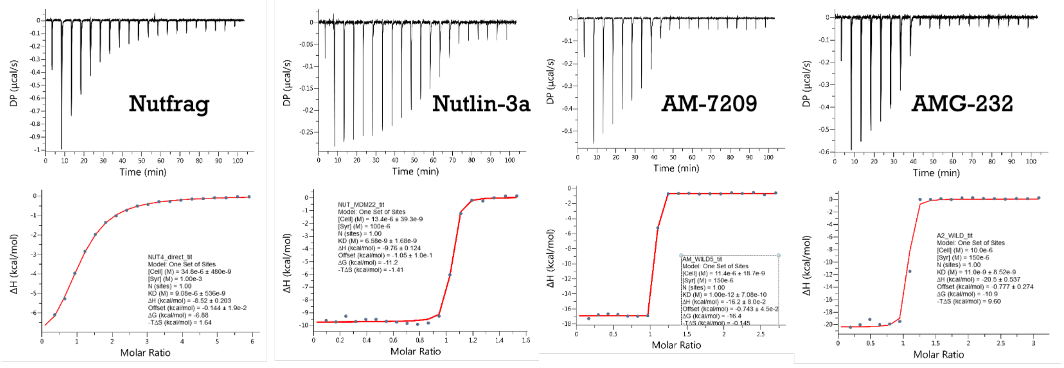


FIGURE E.1: Representative graphs from ITC titrations with the full variant of MDM2 and compounds AM-7209, AMG-232, Nutlin-3a, Nutfrag.

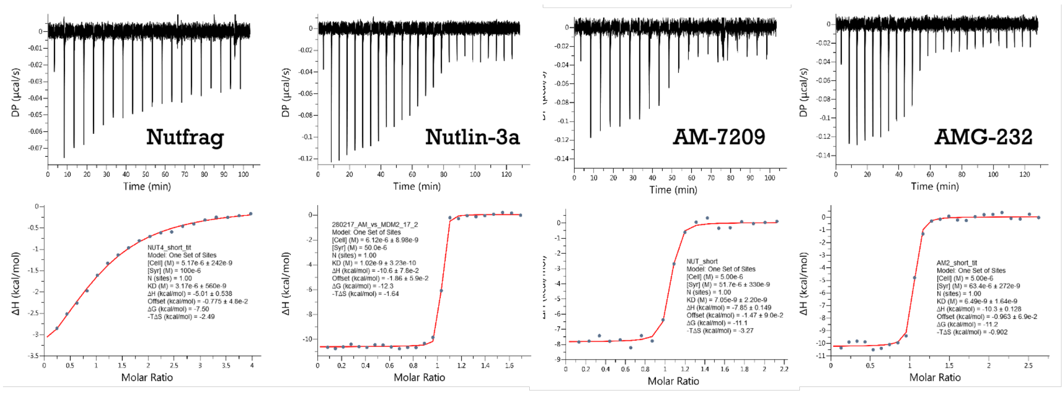


FIGURE E.2: Representative graphs from ITC titrations with the lid absent variant of MDM2 and compounds AM-7209, AMG-232, Nutlin-3a, Nutfrag.

E.2 Convergence and window sampling plots from adaptive sampling protocol

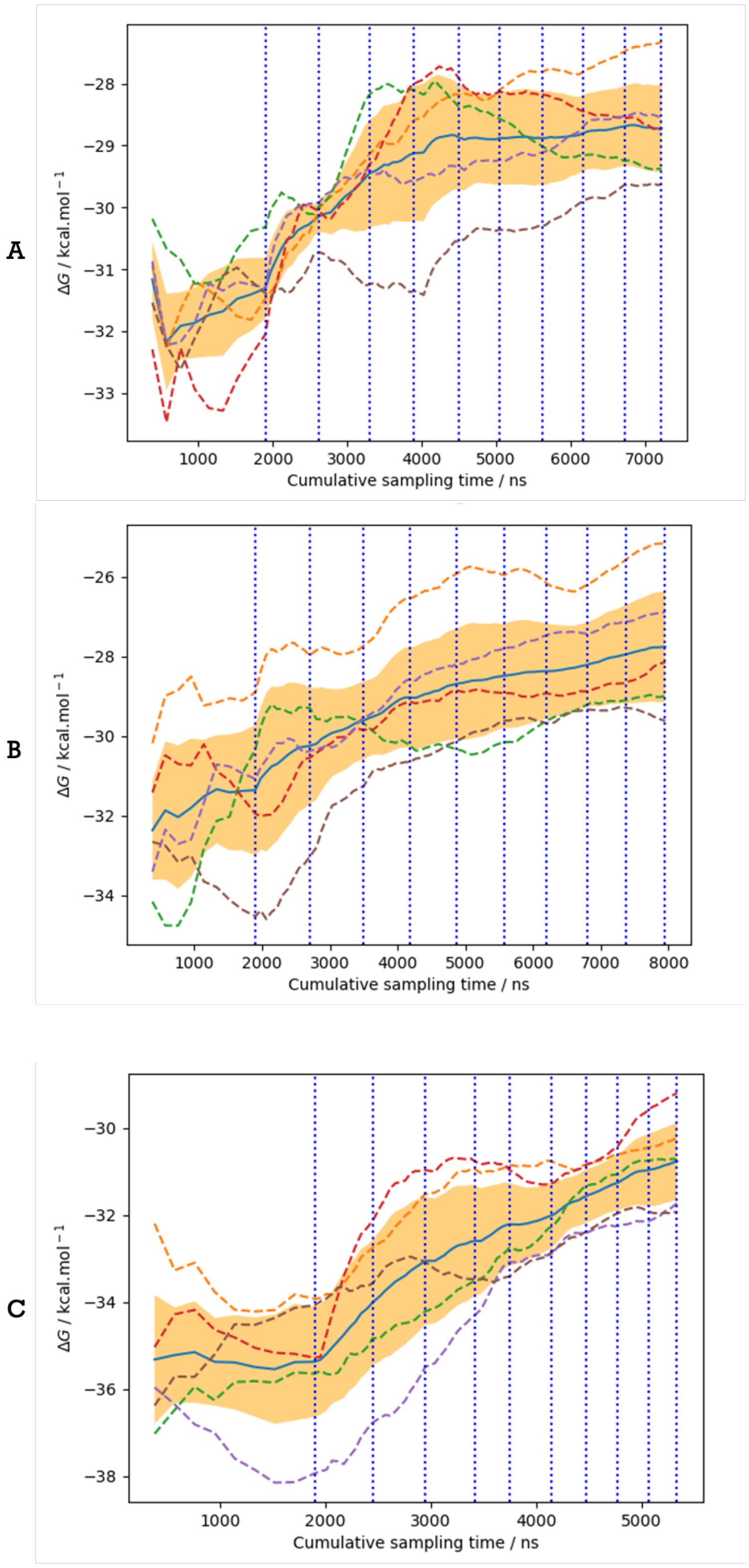


FIGURE E.3: Convergence plots for **A)** MDM2-short/AM-7209, **B)** MDM2-extended-close-lid/AM-7209 and **C)** MDM2-ordered-lid/AM-7209.

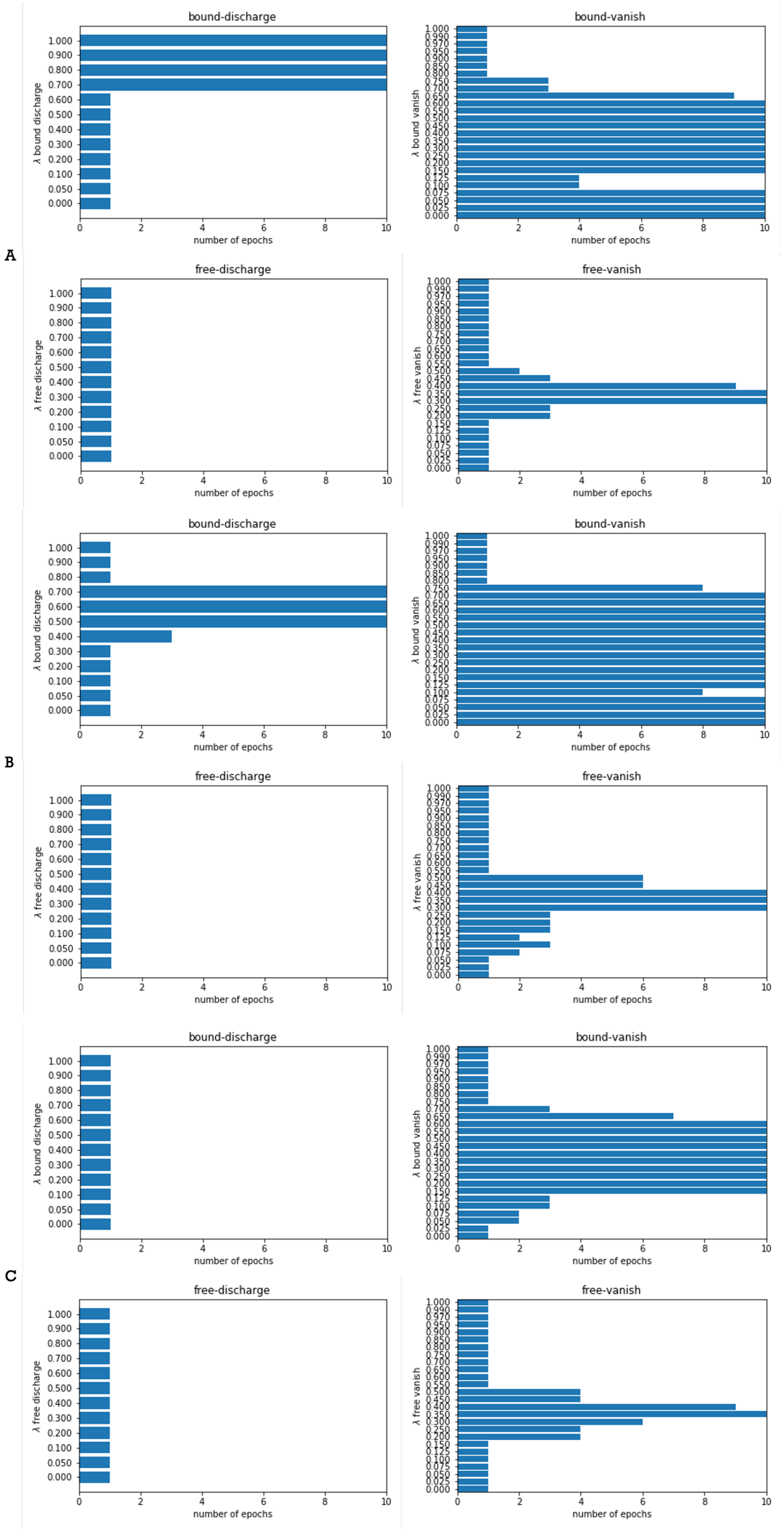


FIGURE E.4: Window sampling plots for **A)** MDM2-short/AM-7209, **B)** MDM2-extended-close-lid/AM-7209 and **C)** MDM2-ordered-lid/AM-7209.

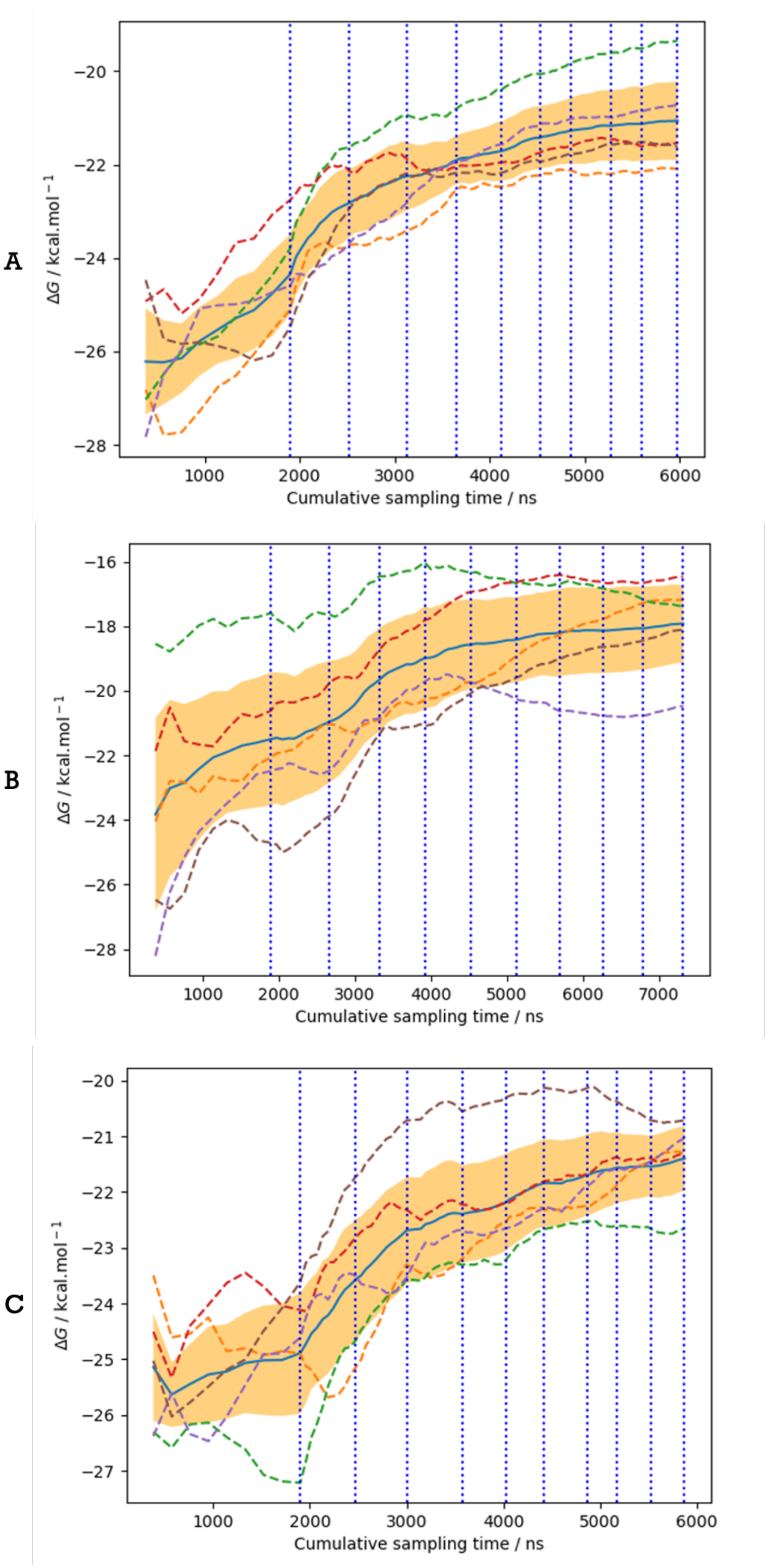


FIGURE E.5: Convergence plots for **A)** MDM2-short/AMG-232, **B)** MDM2-extended-close-lid/AMG-232 and **C)** MDM2-ordered-lid/AMG-232.

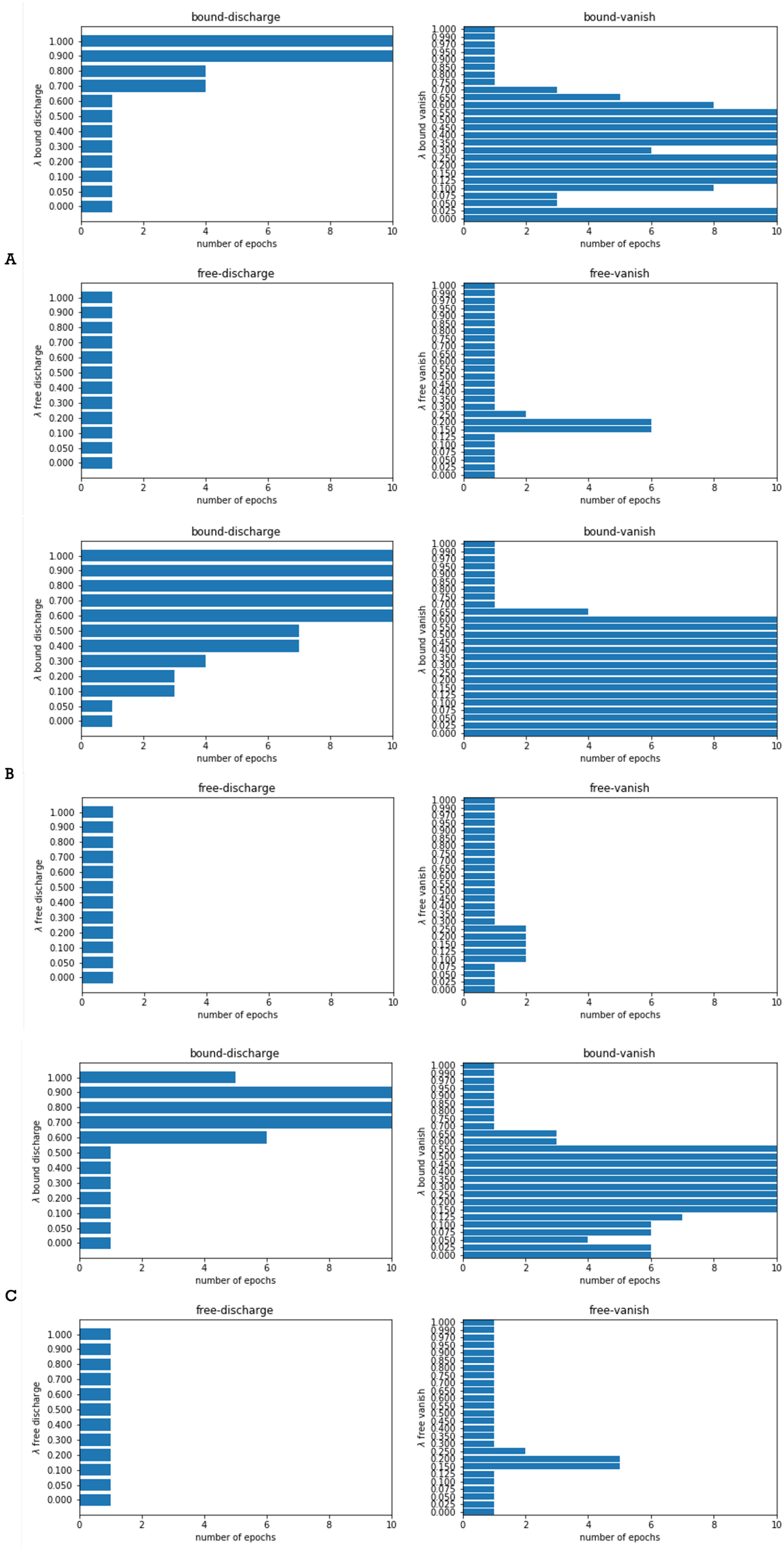


FIGURE E.6: Window sampling plots for **A)** MDM2-short/AMG-232, **B)** MDM2-extended-close-lid/AMG-232 and **C)** MDM2-ordered-lid/AMG-232.

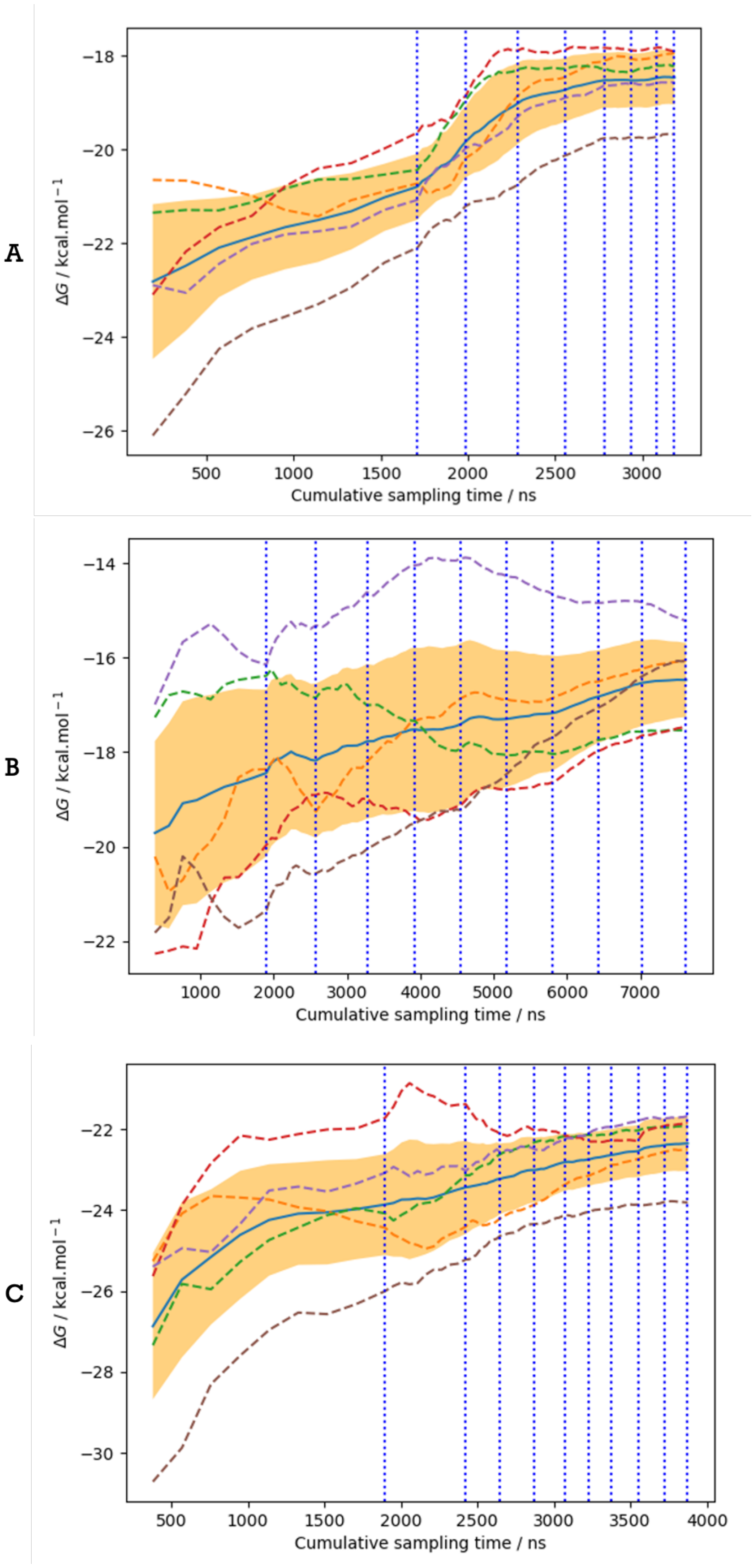


FIGURE E.7: Convergence plots for **A)** MDM2-short/Pip-2, **B)** MDM2-extended-close-lid/Pip-2 and **C)** MDM2-ordered-lid/Pip-2.

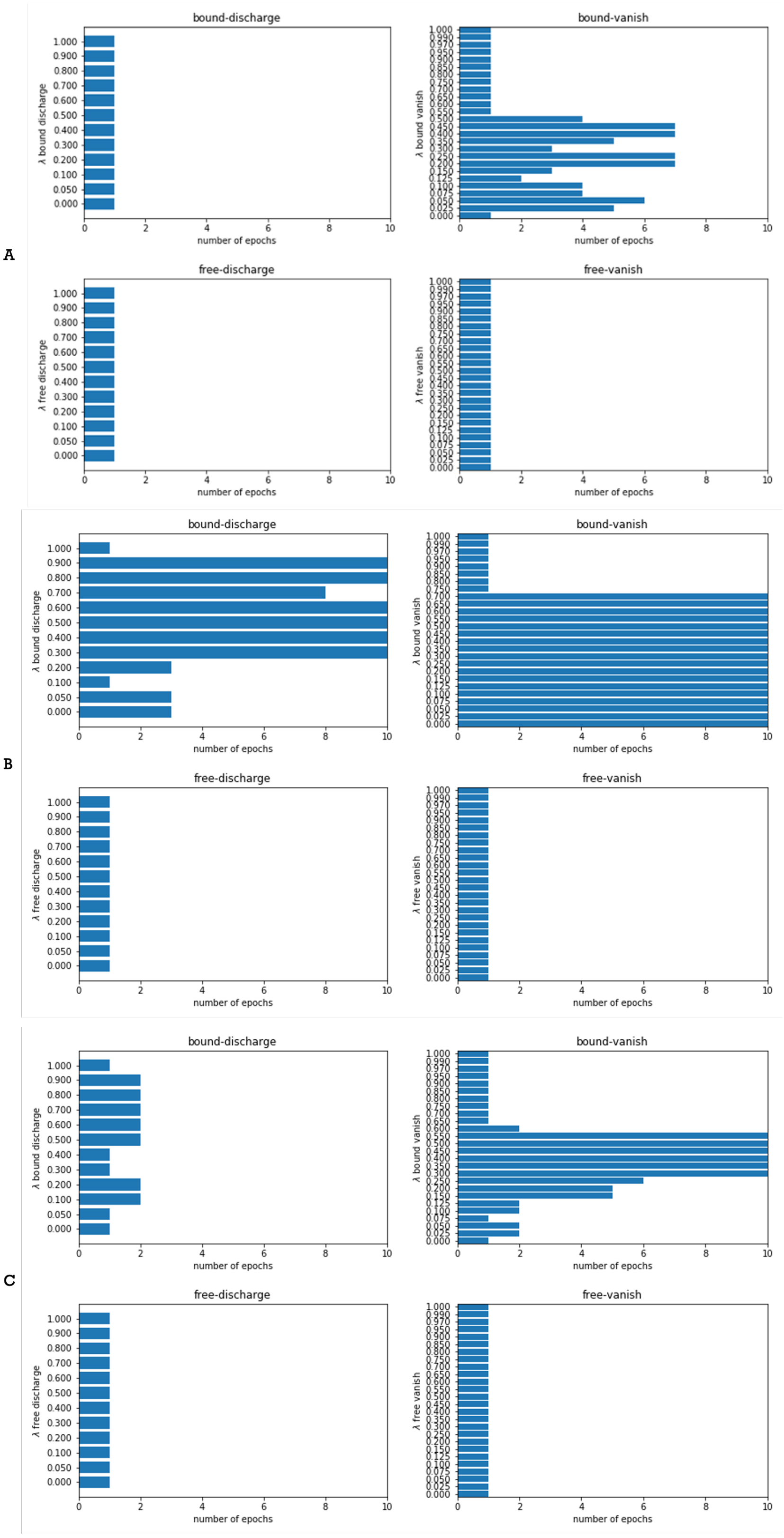


FIGURE E.8: Window sampling plots for **A)** MDM2-short/Pip-2, **B)** MDM2-extended-close-lid/Pip-2 and **C)** MDM2-ordered-lid/Pip-2.

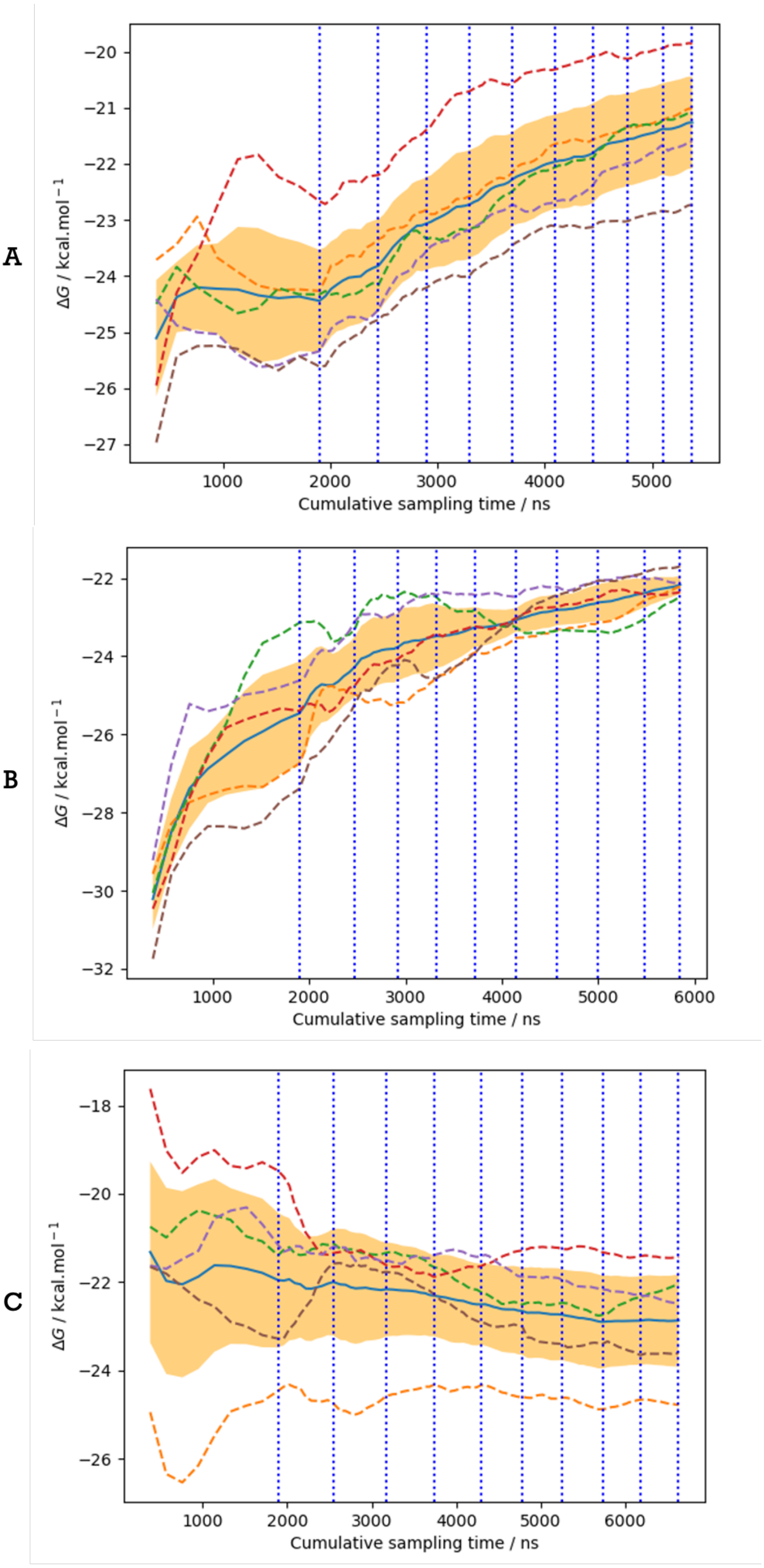


FIGURE E.9: Convergence plots for **A)** MDM2-short/Nutlin-3a, **B)** MDM2-extended-close-lid/Nutlin-3a and **C)** MDM2-ordered-lid/Nutlin-3a.

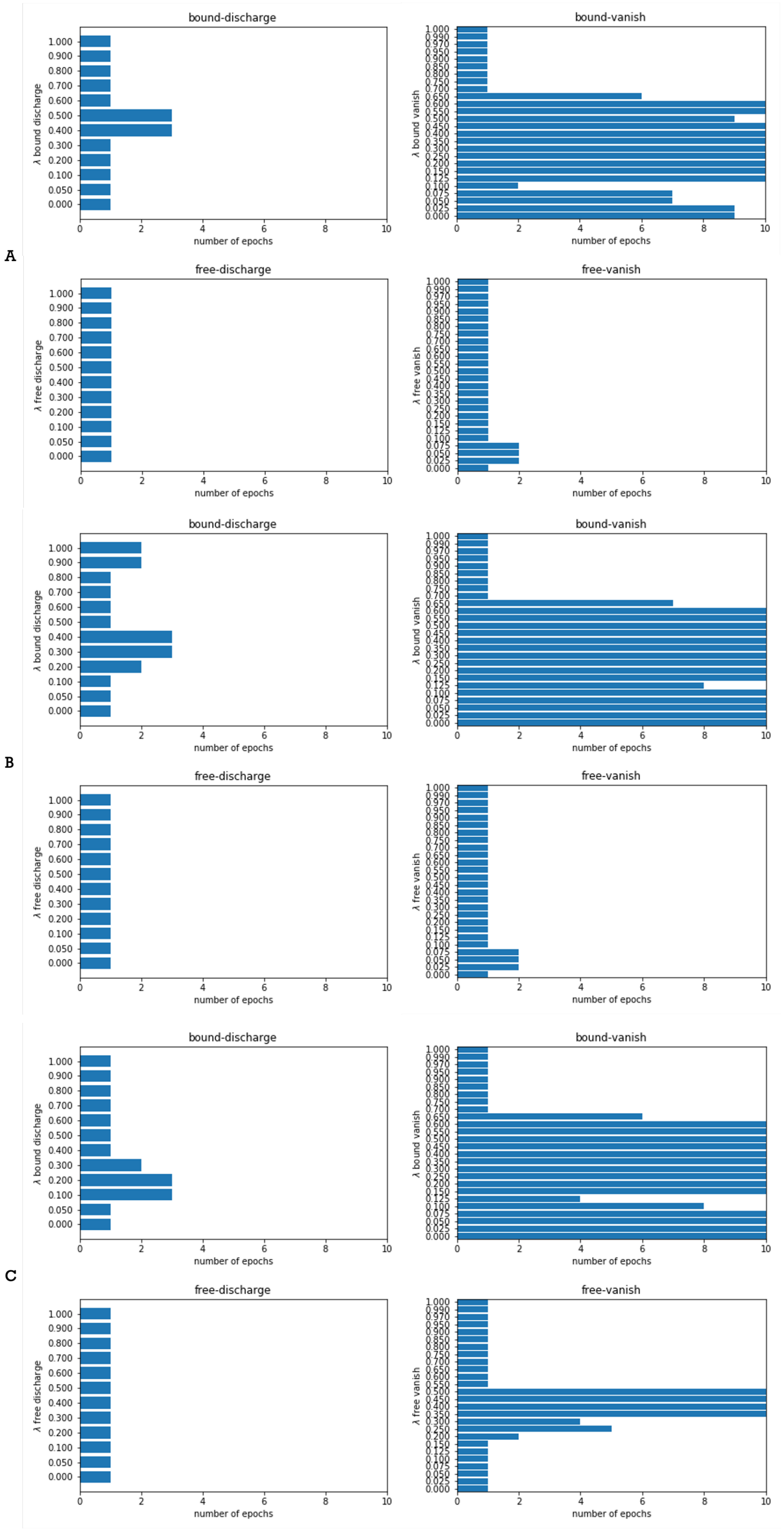


FIGURE E.10: Window sampling plots for **A)** MDM2-short/Nutlin-3a, **B)** MDM2-extended-close-lid/Nutlin-3a and **C)** MDM2-ordered-lid/Nutlin-3a.

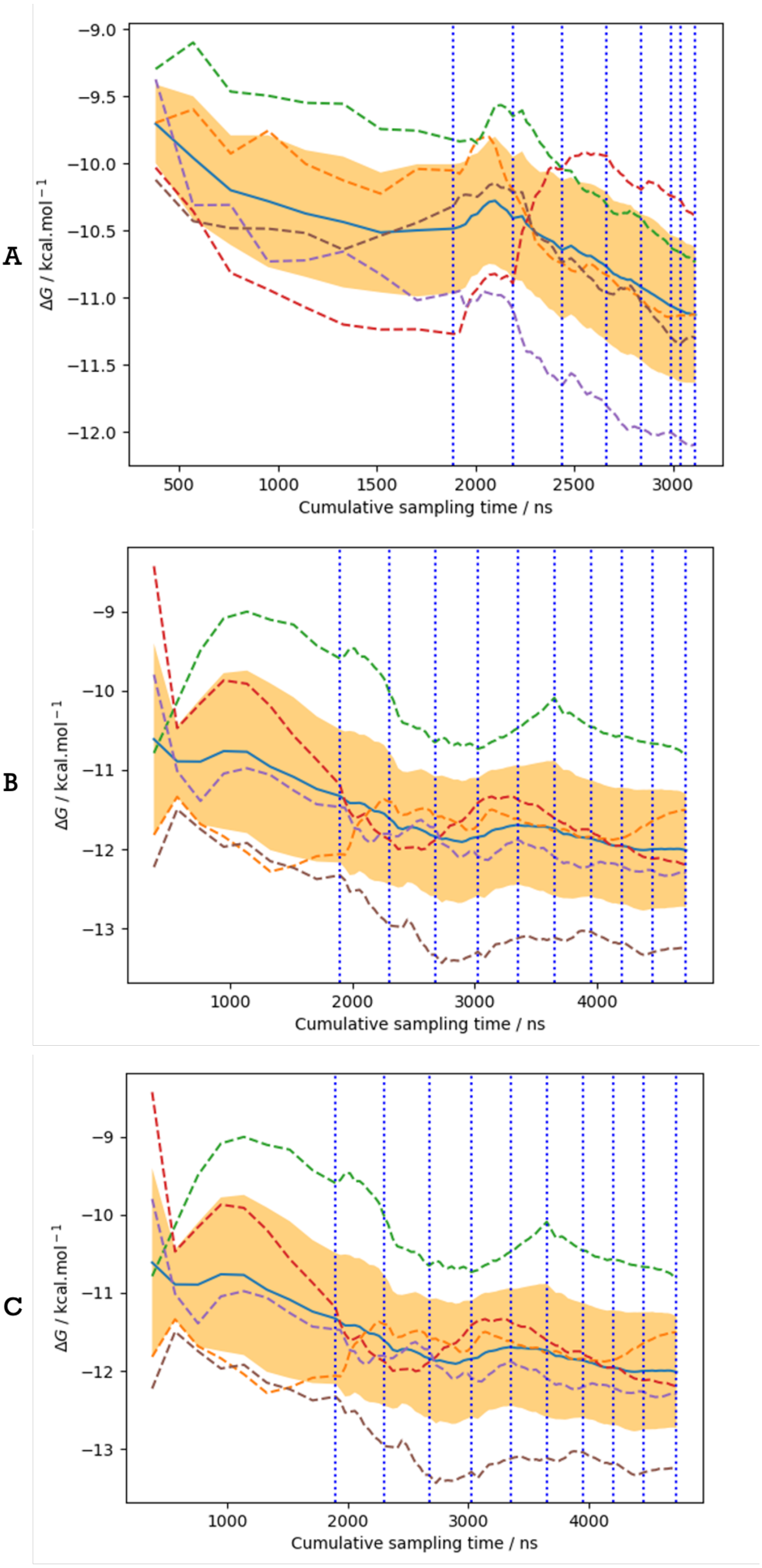


FIGURE E.11: Convergence plots for **A)** MDM2-short/Nutfrag, **B)** MDM2-extended-close-lid/Nutfrag and **C)** MDM2-ordered-lid/Nutfrag.

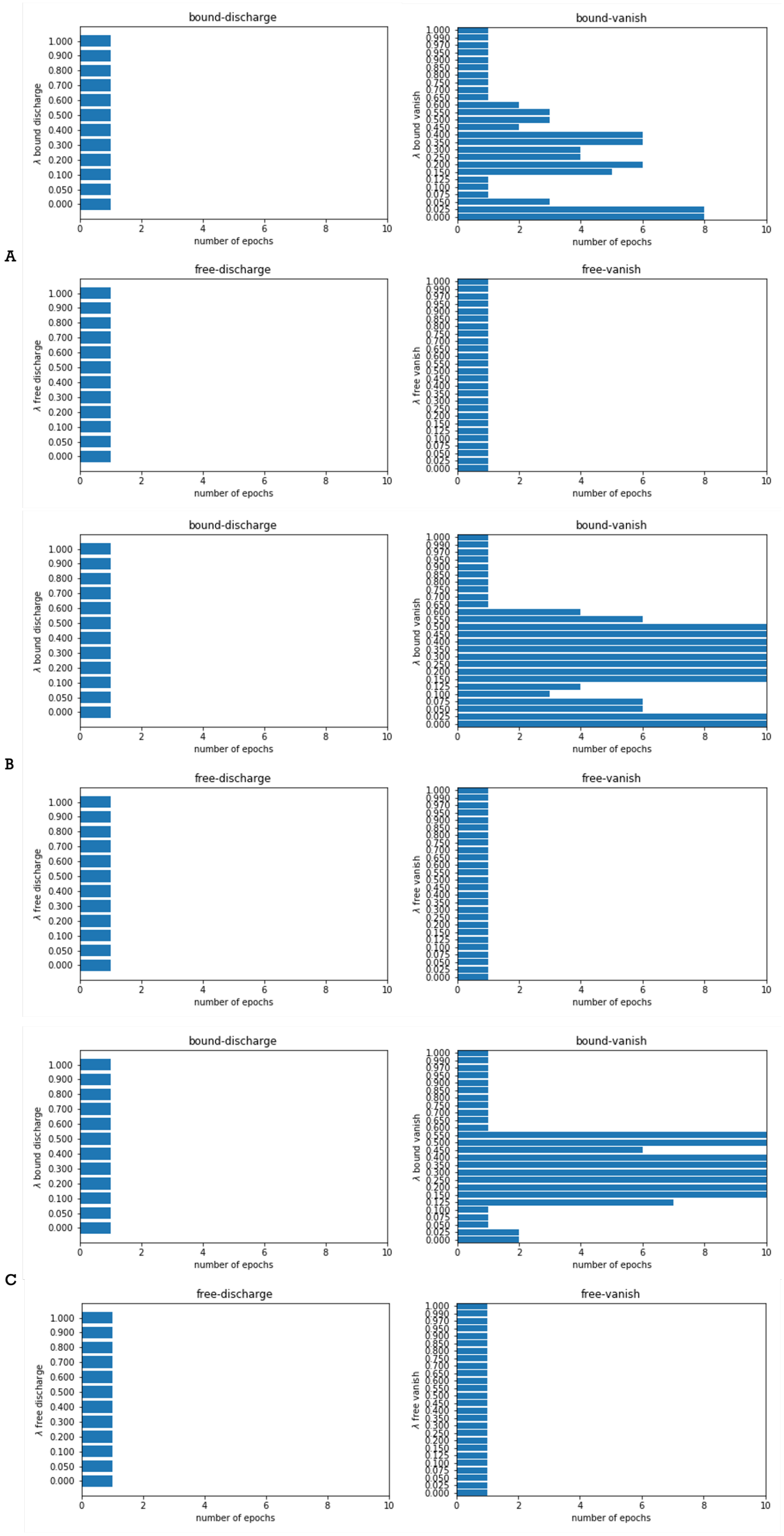


FIGURE E.12: Window sampling plots for **A)** MDM2-short/Nutfrag, **B)** MDM2-extended-close-lid/Nutfrag and **C)** MDM2-ordered-lid/Nutfrag.

E.3 Docking results

Docking studies MDM2					
Ligands	$\Delta G_{MDM2-short}$	$\Delta G_{MDM2-long-flexible}$	$\Delta G_{MDM2-long-ordered}$	$\Delta\Delta G_{ordered/short}$	$\Delta\Delta G_{flexible/short}$
Nutfrag	-4.79 ± 0.01	-8.35 ± 0.01	-6.03 ± 0.02	-1.24 ± 0.01	-3.56 ± 0.01
Nutlin-3a	-10.91 ± 0.05	-8.38 ± 0.13	-7.03 ± 0.03	3.88 ± 0.04	2.53 ± 0.09
Pip-2	-8.00 ± 0.02	-9.47 ± 0.21	-7.67 ± 0.11	0.33 ± 0.06	-1.47 ± 0.11
AMG-232	-8.50 ± -0.08	-9.97 ± 0.14	-9.54 ± 0.14	-1.31 ± 0.13	-1.47 ± 0.11
AM-7209	-10.00 ± 0.10	-14.07 ± 0.15	-10.94 ± 0.04	-0.94 ± 0.07	-4.07 ± 0.13

TABLE E.1: Docking studies for the MDM2 dataset. Energies are reported in kcal mol^{-1}.

Bibliography

[1] B. E. Blass. “Chapter 2 - The Drug Discovery Process: From Ancient Times to the Present Day”. en. *Basic Principles of Drug Discovery and Development*. Ed. by B. E. Blass. Boston: Academic Press, 2015, 35.

[2] J. Drews. en. *Science* **287**. Publisher: American Association for the Advancement of Science Section: Special Reviews, 1960 (2000).

[3] T. A. Ban. *Dialogues in Clinical Neuroscience* **8**, 335 (2006).

[4] M. Golin. en. *Journal of the American Medical Association* **165**, 2084 (1957).

[5] C. Krishnamurti and S. C. Rao. *Indian Journal of Anaesthesia* **60**, 861 (2016).

[6] A. Fleming. *British journal of experimental pathology* **10**, 226 (1929).

[7] A. C. A. Roque, ed. *Ligand-Macromolecular Interactions in Drug Discovery: Methods and Protocols*. en. Methods in Molecular Biology. Humana Press, 2010.

[8] N. Metropolis *et al. The Journal of Chemical Physics* **21**. Publisher: American Institute of Physics, 1087 (1953).

[9] J. A. McCammon, B. R. Gelin, and M. Karplus. en. *Nature* **267**. Number: 5612 Publisher: Nature Publishing Group, 585 (1977).

[10] J. A. Anderson, C. D. Lorenz, and A. Travesset. en. *Journal of Computational Physics* **227**, 5342 (2008).

[11] M. Dickson and J. P. Gagnon. en. *Discovery Medicine* **4**, 172 (2009).

[12] J. Woodcock and R. Woosley. eng. *Annual Review of Medicine* **59**, 1 (2008).

[13] A. C. Anderson. en. *Chemistry & Biology* **10**, 787 (2003).

[14] M. Batool, B. Ahmad, and S. Choi. *International Journal of Molecular Sciences* **20** (2019).

[15] M. C. Peitsch and N. Guex. "Large-Scale Comparative Protein Modelling". en. *Proteome Research: New Frontiers in Functional Genomics*. Ed. by M. R. Wilkins *et al.* Principles and Practice. Berlin, Heidelberg: Springer, 1997, 177.

[16] C. M. Song, S. J. Lim, and J. C. Tong. eng. *Briefings in Bioinformatics* **10**, 579 (2009).

[17] A. Lavecchia and C. Di Giovanni. eng. *Current Medicinal Chemistry* **20**, 2839 (2013).

[18] W. L. Jorgensen. en. *Science* **303**. Publisher: American Association for the Advancement of Science Section: Special Reviews, 1813 (2004).

[19] J. Michel. en. *Physical Chemistry Chemical Physics* **16**. Publisher: The Royal Society of Chemistry, 4465 (2014).

[20] S. M. Larson *et al. arXiv:0901.0866 [physics, q-bio]*. arXiv: 0901.0866 (2009).

[21] D. E. Shaw *et al.* "Millisecond-scale molecular dynamics simulations on Anton". en. *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis - SC '09*. Portland, Oregon: ACM Press, 2009, 1.

[22] K. J. Kohlhoff *et al. Nature chemistry* **6**, 15 (2014).

[23] J. Michel, N. Foloppe, and J. W. Essex. en. *Molecular Informatics* **29**, 570 (2010).

[24] X. Du *et al.* en. *International Journal of Molecular Sciences* **17**. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, 144 (2016).

[25] E. Fischer. *Berichte der deutschen chemischen Gesellschaft* **27**. Publisher: John Wiley & Sons, Ltd, 2985 (1894).

[26] D. E. Koshland. *Proceedings of the National Academy of Sciences of the United States of America* **44**, 98 (1958).

[27] D. W. Miller and K. A. Dill. *Protein Science : A Publication of the Protein Society* **6**, 2166 (1997).

[28] P. Robustelli, S. Piana, and D. E. Shaw. *Journal of the American Chemical Society* **142**. Publisher: American Chemical Society, 11092 (2020).

[29] P. Santofimia-Castaño *et al.* en. *Cellular and Molecular Life Sciences* **77**, 1695 (2020).

[30] L. Boltzmann. en. *Nature* **51**. Number: 1322 Publisher: Nature Publishing Group, 413 (1895).

[31] L. Boltzmann. *Lectures on Gas Theory*. en. Courier Corporation, 1995.

[32] D. A. McQuarrie and M. D. A. *Statistical Mechanics*. en. Google-Books-ID: PANRAAAAMAAJ. Harper & Row, 1975.

[33] M. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. en. Google-Books-ID: Lo3Jqc0pgrcC. OUP Oxford, 2010.

[34] G. D. Birkhoff. *Proceedings of the National Academy of Sciences of the United States of America* **17**, 656 (1931).

[35] M. Born and R. Oppenheimer. en. *Annalen der Physik* **389**, 457 (1927).

[36] D. J. Wales. *Energy landscapes*. Cambridge molecular science. Cambridge, UK ; New York: Cambridge University Press, 2003.

[37] *Lennard-Jones potential @ Chemistry Dictionary & Glossary*.

[38] D. Frenkel and B. Smit. *Understanding molecular simulation from algorithms to applications*. English. OCLC: 890552742. San Diego [u.a.: Academic Press, 2002.

[39] V. S. Inakollu *et al.* en. *Current Opinion in Structural Biology*. Theory and Simulation Macromolecular Assemblies**61**, 182 (2020).

[40] *Torsional Angle Interactions*.

[41] J. A. White *et al.* en. *Physica A: Statistical Mechanics and its Applications* **387**, 6705 (2008).

[42] R. S. Katiyar and P. K. Jha. en. *WIREs Computational Molecular Science* **8**, e1358 (2018).

[43] J. Kolafa and J. W. Perram. *Molecular Simulation* **9**. Publisher: Taylor & Francis _: https://doi.org/10.1080/08927029208049126, 351 (1992).

[44] I. Fukuda and H. Nakamura. en. *Biophysical Reviews* **4**, 161 (2012).

[45] C. J. Cramer. *Essentials of computational chemistry: theories and models*. 2nd ed. Chichester, West Sussex, England ; Hoboken, NJ: Wiley, 2004.

[46] D. Fincham. en. *Computer Physics Communications* **40**, 263 (1986).

[47] W. E and D. Li. en. *Communications on Pure and Applied Mathematics* **61**, 96 (2008).

[48] H. C. Andersen. *The Journal of Chemical Physics* **72**. Publisher: American Institute of Physics, 2384 (1980).

[49] R. W. Zwanzig. *The Journal of Chemical Physics* **23**. Publisher: American Institute of Physics, 1915 (1955).

[50] M. R. Shirts and J. D. Chodera. *The Journal of Chemical Physics* **129** (2008).

[51] M. J. Mitchell and J. A. McCammon. en. *Journal of Computational Chemistry* **12**, 271 (1991).

[52] J. Michel and J. W. Essex. eng. *Journal of Computer-Aided Molecular Design* **24**, 639 (2010).

[53] W. L. Jorgensen *et al.* en. *The Journal of Chemical Physics* **89**, 3742 (1988).

[54] M. K. Gilson *et al. Biophysical Journal* **72**, 1047 (1997).

[55] G. Pérez-Hernández and F. Noé. *Journal of Chemical Theory and Computation* **12**. Publisher: American Chemical Society, 6118 (2016).

[56] J. A. Hartigan. *Clustering Algorithms*. en. Google-Books-ID: cDnvAAAAMAAJ. Wiley, 1975.

[57] L.-T. Da *et al.* eng. *Advances in Experimental Medicine and Biology* **805**, 29 (2014).

[58] A. K. Jain. en. *Pattern Recognition Letters*. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR)**31**, 651 (2010).

[59] J.-H. Prinz *et al. The Journal of Chemical Physics* **134**. Publisher: American Institute of Physics, 174105 (2011).

[60] B. Trendelkamp-Schroer *et al. The Journal of Chemical Physics* **143**. Publisher: American Institute of Physics, 174101 (2015).

[61] W. C. Swope, J. W. Pitera, and F. Suits. *The Journal of Physical Chemistry B* **108**. Publisher: American Chemical Society, 6571 (2004).

[62] G. R. Bowman, V. S. Pande, and F. Noé, eds. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. Vol. 797. Advances in Experimental Medicine and Biology. Dordrecht: Springer Netherlands, 2014.

[63] F. Noé *et al.* en. *Proceedings of the National Academy of Sciences* **106**, 19011 (2009).

[64] S. Röblitz and M. Weber. en. *Advances in Data Analysis and Classification* **7**, 147 (2013).

[65] P. Deuflhard and M. Weber. en. *Linear Algebra and its Applications* **398**, 161 (2005).

[66] S. Kube and M. Weber. *The Journal of Chemical Physics* **126**. Publisher: American Institute of Physics, 024103 (2007).

[67] P. Deuflhard *et al.* en. *Linear Algebra and its Applications* **315**, 39 (2000).

[68] F. Noé and S. Fischer. en. *Current Opinion in Structural Biology*. Theory and simulation / Macromolecular assemblages**18**, 154 (2008).

[69] N. Singhal and V. S. Pande. *The Journal of Chemical Physics* **123**. Publisher: American Institute of Physics, 204909 (2005).

[70] X. HUANG *et al. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 228 (2010).

[71] A. K. Bronowska. en. *Thermodynamics - Interaction Studies - Solids, Liquids and Gases*. Publisher: IntechOpen (2011).

[72] A. Mittermaier and L. E. Kay. eng. *Science (New York, N.Y.)* **312**, 224 (2006).

[73] R. O'Brien, N. Markova, and G. A. Holdgate. "Thermodynamics in Drug Discovery". en. *Applied Biophysics for Drug Discovery*. Section: 2 _: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119099512.ch2. John Wiley & Sons, Ltd, 2017, 7.

[74] M. M. Pierce, C. S. Raman, and B. T. Nall. eng. *Methods (San Diego, Calif.)* **19**, 213 (1999).

[75] K. Narayan and S. S. Carroll. "SPR Screening". en. *Applied Biophysics for Drug Discovery*. Section: 6 _: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ John Wiley & Sons, Ltd, 2017, 93.

[76] S. G. Patching. en. *Biochimica et Biophysica Acta (BBA) - Biomembranes*. Structural and biophysical characterisation of membrane protein-ligand binding**1838**, 43 (2014).

[77] W. A. Lea and A. Simeonov. *Expert opinion on drug discovery* **6**, 17 (2011).

[78] A. M. Rossi and C. W. Taylor. *Nature protocols* **6**, 365 (2011).

[79] Y. Cheng and W. H. Prusoff. eng. *Biochemical Pharmacology* **22**, 3099 (1973).

[80] P. J. Munson and D. Rodbard. eng. *Journal of Receptor Research* **8**, 533 (1988).

[81] M. Aldeghi *et al.* en. *Chemical Science* **7**. Publisher: The Royal Society of Chemistry, 207 (2015).

[82] G. Fischer, H. Bang, and C. Mech. ger. *Biomedica Biochimica Acta* **43**, 1101 (1984).

[83] N. Takahashi, T. Hayano, and M. Suzuki. eng. *Nature* **337**, 473 (1989).

[84] S. F. Göthel and M. A. Marahiel. eng. *Cellular and molecular life sciences: CMLS* **55**, 423 (1999).

[85] P. A. M. Schmidpeter and F. X. Schmid. en. *Journal of Molecular Biology*. Molecular Chaperones and Protein Quality Control (Part I)**427**, 1609 (2015).

[86] T. Kiefhaber *et al. Biochemistry* **29**. Publisher: American Chemical Society, 3061 (1990).

[87] P. Wang and J. Heitman. *Genome Biology* **6**, 226 (2005).

[88] T. L. Davis *et al.* en. *PLoS Biology* **8**,ed. by G. A. Petsko, e1000439 (2010).

[89] J. Liu *et al.* eng. *Cell* **66**, 807 (1991).

[90] D. S. Horowitz *et al. The EMBO Journal* **21**, 470 (2002).

[91] L. A. Gaither *et al. Virology* **397**, 43 (2010).

[92] P. Gallay and Lin. en. *Drug Design, Development and Therapy*, 105 (2013).

[93] P. Nigro, G. Pompilio, and M. C. Capogrossi. eng. *Cell Death & Disease* **4**, e888 (2013).

[94] J. Colgan *et al.* eng. *Journal of Immunology (Baltimore, Md.: 1950)* **174**, 6030 (2005).

[95] H. Hoffmann and C. Schiene-Fischer. eng. *Biological Chemistry* **395**, 721 (2014).

[96] B. A. Howard *et al.* eng. *Lung Cancer (Amsterdam, Netherlands)* **46**, 313 (2004).

[97] B. A. Howard *et al.* eng. *Cancer Research* **65**, 8853 (2005).

[98] M. Li *et al.* eng. *Cancer* **106**, 2284 (2006).

[99] S. O. Lim *et al.* eng. *Biochemical and Biophysical Research Communications* **291**, 1031 (2002).

[100] M. Al-Ghoul *et al.* eng. *Journal of Proteome Research* **7**, 4107 (2008).

[101] X. Han *et al.* eng. *Oncology Reports* **23**, 1053 (2010).

[102] Y.-J. Qi *et al.* eng. *Journal of Cellular Biochemistry* **104**, 1625 (2008).

[103] K. J. Choi *et al.* eng. *Cancer Research* **67**, 3654 (2007).

[104] C. Melle *et al.* eng. *International Journal of Molecular Medicine* **16**, 11 (2005).

[105] J. Lee and S. S. Kim. *Journal of Experimental & Clinical Cancer Research : CR* **29**, 97 (2010).

[106] T. Dorfman *et al.* eng. *Journal of Virology* **71**, 7110 (1997).

[107] K. Zander *et al.* eng. *The Journal of Biological Chemistry* **278**, 43202 (2003).

[108] M. Qi and C. Aiken. eng. *Virology* **373**, 287 (2008).

[109] X. Hanoulle *et al. The Journal of Biological Chemistry* **284**, 13589 (2009).

[110] F. Fernandes, I.-u. H. Ansari, and R. Striker. en. *PLOS ONE* **5**, e9815 (2010).

[111] U. Chatterji *et al.* eng. *The Journal of Biological Chemistry* **284**, 16998 (2009).

[112] H. Tang. *Viruses* **2**, 1621 (2010).

[113] N. V. Naoumov. *Journal of Hepatology* **61**, 1166 (2014).

[114] H. Du *et al.* eng. *Biochimica Et Biophysica Acta* **1842**, 2517 (2014).

[115] A. Rasola and F. Chiara. English. *Frontiers in Oncology* **3** (2013).

[116] A. Linkermann and D. R. Green. eng. *The New England Journal of Medicine* **370**, 455 (2014).

[117] O. N. Tucker and N. Heaton. eng. *Current Opinion in Critical Care* **11**, 150 (2005).

[118] H. Rehman *et al. The Journal of pharmacology and experimental therapeutics* **327**, 699 (2008).

[119] Y. Wei *et al. World Journal of Gastroenterology : WJG* **14**, 193 (2008).

[120] D. Pessayre and B. Fromenty. English. *Journal of Hepatology* **42**. Publisher: Elsevier, 928 (2005).

[121] Z. M. Younossi. en. *Clinical Liver Disease* **11**, 92 (2018).

[122] *The NASH Education Program^{TM} | Learn more about the initiative*. en-US. Library Catalog: www.the-nash-education-program.com.

[123] *The Emerging NASH Crisis | Liver Disease | TrialSite News*. en-US. 2019.

[124] X. Wang *et al.* en. *Hepatology* **68**, 62 (2018).

[125] S. J. Pandol *et al.* English. *Gastroenterology* **132**. Publisher: Elsevier, 1127 (2007).

[126] M. S. Petrov *et al.* English. *Gastroenterology* **139**. Publisher: Elsevier, 813 (2010).

[127] R. Mukherjee *et al. Gut* **65**, 1333 (2016).

[128] J. W. Elrod and J. D. Molkentin. *Circulation Journal* **77**, 1111 (2013).

[129] E. R. Shore *et al. Journal of Medicinal Chemistry* **59**. Publisher: American Chemical Society, 2596 (2016).

[130] H. Du *et al.* en. *Nature Medicine* **14**. Number: 10 Publisher: Nature Publishing Group, 1097 (2008).

[131] H. Svarstad, H. C. Bugge, and S. S. Dhillion. en. *Biodiversity & Conservation* **9**, 1521 (2000).

[132] H. F. Stähelin. en. *Experientia* **52**, 5 (1996).

[133] R. G. Ptak *et al.* en. *Antimicrobial Agents and Chemotherapy* **52**, 1302 (2008).

[134] L. Coelmont *et al.* en. *PLOS ONE* **5**. Publisher: Public Library of Science, e13687 (2010).

[135] R. Flisiak *et al. Expert Opinion on Investigational Drugs* **21**, 375 (2012).

[136] Z. K. Sweeney, J. Fu, and B. Wiedmann. *Journal of Medicinal Chemistry* **57**. Publisher: American Chemical Society, 7145 (2014).

[137] L. J. Anderson *et al. Virology Journal* **8**, 329 (2011).

[138] J. E. Mathy *et al.* en. *Antimicrobial Agents and Chemotherapy* **52**. Publisher: American Society for Microbiology Journals Section: ANTIVIRAL AGENTS, 3267 (2008).

[139] H. Harada *et al.* en. *Blood* **108**. Publisher: American Society of Hematology, 1253 (2006).

[140] S. Hopkins and P. Gallay. en. *Viruses* **4**. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, 2558 (2012).

[141] S. Hopkins *et al.* en. *Antimicrobial Agents and Chemotherapy* **54**. Publisher: American Society for Microbiology Journals Section: ANTIVIRAL AGENTS, 660 (2010).

[142] R. Sedrani *et al.* *Journal of the American Chemical Society* **125**. Publisher: American Chemical Society, 3849 (2003).

[143] C. Härtel *et al.* en. *Scandinavian Journal of Immunology* **63**, 26 (2006).

[144] S. N. Immecke *et al.* en. *PLOS ONE* **6**. Publisher: Public Library of Science, e18406 (2011).

[145] J. Kallen *et al.* en. *Journal of Biological Chemistry* **280**. Publisher: American Society for Biochemistry and Molecular Biology, 21965 (2005).

[146] K. Goto *et al.* en. *Cancer Science* **100**, 1943 (2009).

[147] M. A. Gregory *et al.* en. *Antimicrobial Agents and Chemotherapy* **55**. Publisher: American Society for Microbiology Journals Section: Antiviral Agents, 1975 (2011).

[148] R. L. Mackman *et al.* *Journal of Medicinal Chemistry* **61**. Publisher: American Chemical Society, 9473 (2018).

[149] S. Hopkins and P. Gallay. en. *Viruses* **4**, 2558 (2012).

[150] C. J. Dunsmore *et al.* *ChemBioChem* **12**. Publisher: John Wiley & Sons, Ltd, 802 (2011).

[151] J.-F. Guichou *et al.* en. *Journal of Medicinal Chemistry* **49**, 900 (2006).

[152] S. Ni *et al.* en. *Journal of Medicinal Chemistry* **52**, 5295 (2009).

[153] S. Yang *et al.* eng. *Journal of Medicinal Chemistry* **58**, 9546 (2015).

[154] A. Ahmed-Belkacem *et al.* en. *Nature Communications* **7**. Number: 1 Publisher: Nature Publishing Group, 1 (2016).

[155] P. Taylor *et al.* en. *Progress in Biophysics and Molecular Biology* **67**, 155 (1997).

[156] A. D. Simone *et al.* en. *Chemical Science* **10**, 542 (2019).

[157] J. Zheng *et al.* en. *Cancer Research* **68**. Publisher: American Association for Cancer Research Section: Cell, Tumor, and Stem Cell Biology, 7769 (2008).

[158] P Tosco *et al.* en, 1 ().

[159] J. J. Irwin and B. K. Shoichet. *Journal of chemical information and modeling* **45**, 177 (2005).

[160] A. Gaulton *et al. Nucleic Acids Research* **40**, D1100 (2012).

[161] C. R. Groom *et al. Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials* **72**, 171 (2016).

[162] T. Cheeseright *et al.* eng. *Journal of Chemical Information and Modeling* **46**, 665 (2006).

[163] M. Slater and A. Vinter. "The XED Force Field and Spark". en. *Scaffold Hopping in Medicinal Chemistry*. John Wiley & Sons, Ltd, 2013, 195.

[164] J. A. Grant, M. A. Gallardo, and B. T. Pickup. en. *Journal of Computational Chemistry* **17**, 1653 (1996).

[165] O. V. Stroganov *et al.* eng. *Journal of Chemical Information and Modeling* **48**, 2371 (2008).

[166] *RDKit*.

[167] H. H. Loeffler, J. Michel, and C. Woods. *Journal of Chemical Information and Modeling* **55**. Publisher: American Chemical Society, 2485 (2015).

[168] D. Case *et al.* Publisher: University of California, San Francisco (2017).

[169] J. A. Maier *et al. Journal of chemical theory and computation* **11**, 3696 (2015).

[170] J. Wang *et al.* en. *Journal of Computational Chemistry* **25**, 1157 (2004).

[171] J. Wang *et al.* en. *Journal of Molecular Graphics and Modelling* **25**, 247 (2006).

[172] A. Jakalian, D. B. Jack, and C. I. Bayly. fr. *Journal of Computational Chemistry* **23**, 1623 (2002).

[173] W. L. Jorgensen *et al. The Journal of Chemical Physics* **79**. Publisher: American Institute of Physics, 926 (1983).

[174] A. S.J. S. Mey, J. J. Jiménez, and J. Michel. en. *Journal of Computer-Aided Molecular Design* **32**, 199 (2018).

[175] A. A. Hagberg, D. A. Schult, and P. J. Swart. en, 5 (2008).

[176] *ChemAxon - Software Solutions and Services for Chemistry & Biology*.

[177] U. Grädler *et al.* en. *Bioorganic & Medicinal Chemistry Letters* **29**, 126717 (2019).

[178] Y. Deng and B. Roux. *Journal of Chemical Theory and Computation* **2**. Publisher: American Chemical Society, 1255 (2006).

[179] C.-E. Chang and M. K. Gilson. *Journal of the American Chemical Society* **126**. Publisher: American Chemical Society, 13156 (2004).

[180] G. Calabrò *et al. The Journal of Physical Chemistry B* **120**. Publisher: American Chemical Society, 5340 (2016).

[181] L. Wang *et al. Journal of the American Chemical Society* **137**. Publisher: American Chemical Society, 2695 (2015).

[182] S. Bosisio, A. S.J. S. Mey, and J. Michel. en. *Journal of Computer-Aided Molecular Design* **31**, 61 (2017).

[183] S. Bosisio, A. S.J. S. Mey, and J. Michel. en. *Journal of Computer-Aided Molecular Design* **30**, 1101 (2016).

[184] C. C. Bannan *et al.* en. *Journal of Computer-Aided Molecular Design* **30**, 927 (2016).

[185] A. Rodil *et al.* en. *Chemical Science* **9**. Publisher: The Royal Society of Chemistry, 3023 (2018).

[186] J. D. Chodera *et al. Current opinion in structural biology* **21**, 150 (2011).

[187] I.-J. Chen and N. Foloppe. en. *Drug Development Research* **72**, 85 (2011).

[188] M. Souaille and B. Roux. en. *Computer Physics Communications* **135**, 40 (2001).

[189] H. Li, M. Fajer, and W. Yang. *The Journal of Chemical Physics* **126**. Publisher: American Institute of Physics, 024106 (2007).

[190] T. A. Halgren and W. Damm. en. *Current Opinion in Structural Biology* **11**, 236 (2001).

[191] M. A. Kastenholz and P. H. Hünenberger. *The Journal of Physical Chemistry B* **108**. Publisher: American Chemical Society, 774 (2004).

[192] M. M. Reif and C. Oostenbrink. *Journal of Computational Chemistry* **35**, 227 (2014).

[193] G. J. Rocklin *et al.* *The Journal of Chemical Physics* **139** (2013).

[194] A. S.J. S. Mey *et al.* en. *Bioorganic & Medicinal Chemistry*. Advances in Computational and Medicinal Chemistry**24**, 4890 (2016).

[195] Z. Gaieb *et al.* *Journal of computer-aided molecular design* **32**, 1 (2018).

[196] A. Nicholls *et al.* *Journal of Medicinal Chemistry* **51**. Publisher: American Chemical Society, 769 (2008).

[197] D. L. Mobley *et al.* *Journal of computer-aided molecular design* **26**, 551 (2012).

[198] T. S. Peat *et al.* *Journal of computer-aided molecular design* **28**, 347 (2014).

[199] D. L. Mobley and M. K. Gilson. *Annual Review of Biophysics* **46**, 531 (2017).

[200] P. Mikulskis *et al.* en. *Journal of Computer-Aided Molecular Design* **28**, 375 (2014).

[201] G. König *et al.* *Journal of computer-aided molecular design* **28**, 245 (2014).

[202] J. I. Monroe and M. R. Shirts. en. *Journal of Computer-Aided Molecular Design* **28**, 401 (2014).

[203] M. Mezei. en. *The Journal of Chemical Physics* **86**, 7084 (1987).

[204] W. L. Jorgensen and C. Ravimohan. *The Journal of Chemical Physics* **83**. Publisher: American Institute of Physics, 3050 (1985).

[205] B. R. Miller *et al.* *Journal of Chemical Theory and Computation* **8**. Publisher: American Chemical Society, 3314 (2012).

[206] H. Gan, C. J. Benjamin, and B. C. Gibb. *Journal of the American Chemical Society* **133**. Publisher: American Chemical Society, 4770 (2011).

[207] C. L. D. Gibb and B. C. Gibb. *Journal of computer-aided molecular design* **28**, 319 (2014).

[208] M. R. Sullivan *et al.* *Journal of computer-aided molecular design* **31**, 21 (2017).

[209] H. Gan and B. C. Gibb. en. *Chemical Communications* **49**. Publisher: The Royal Society of Chemistry, 1395 (2013).

[210] K. I. Assaf and W. M. Nau. en. *Chemical Society Reviews* **44**. Publisher: The Royal Society of Chemistry, 394 (2014).

[211] F. Biedermann and O. A. Scherman. *The Journal of Physical Chemistry B* **116**. Publisher: American Chemical Society, 2842 (2012).

[212] J. Vázquez *et al.* en. *Chemistry – A European Journal* **20**, 9897 (2014).

[213] S. Liu *et al. Journal of the American Chemical Society* **127**. Publisher: American Chemical Society, 15959 (2005).

[214] World Health Organization. *World Malaria Report 2015*. English. OCLC: 969445251. World Health Organization, 2016.

[215] A. Rizzi *et al.* en. *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, 795005 (2019).

[216] M. R. Shirts *et al. The Journal of Physical Chemistry B* **111**. Publisher: American Chemical Society, 13052 (2007).

[217] M. McGann. *Journal of Chemical Information and Modeling* **51**. Publisher: American Chemical Society, 578 (2011).

[218] M. McGann. en. *Journal of Computer-Aided Molecular Design* **26**, 897 (2012).

[219] B. P. Kelley *et al. Journal of Chemical Information and Modeling* **55**. Publisher: American Chemical Society, 1771 (2015).

[220] *ParmEd — ParmEd documentation*.

[221] C. Woods *et al. Sire molecular simulations framework*. 2016.

[222] P. Eastman *et al. Journal of chemical theory and computation* **9**, 461 (2013).

[223] I. G. Tironi *et al. The Journal of Chemical Physics* **102**. Publisher: American Institute of Physics, 5451 (1995).

[224] J. C. Phillips *et al.* en. *Journal of Computational Chemistry* **26**, 1781 (2005).

[225] M. J. Abraham *et al.* en. *SoftwareX* **1-2**, 19 (2015).

[226] A. Rizzi *et al.* en. *Journal of Computer-Aided Molecular Design* **34**, 601 (2020).

[227] A. Rizzi *et al.* en. *Journal of Computer-Aided Molecular Design* **32**, 937 (2018).

[228] S. Murkli, J. N. McNeill, and L. Isaacs. *Supramolecular Chemistry* **31**. Publisher: Taylor & Francis _: https://doi.org/10.1080/10610278.2018.1516885, 150 (2019).

[229] M. R. Sullivan, W. Yao, and B. C. Gibb. *Supramolecular Chemistry* **31**. Publisher: Taylor & Francis _: https://doi.org/10.1080/10610278.2018.1549327, 184 (2019).

[230] P. E. Wright and H. J. Dyson. en. *Journal of Molecular Biology* **293**, 321 (1999).

[231] M. Sickmeier *et al. Nucleic Acids Research* **35**, D786 (2007).

[232] A. K. Dunker *et al.* eng. *BMC genomics* **9 Suppl 2**, S1 (2008).

[233] V. N. Uversky. en. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1814**, 693 (2011).

[234] A. Schlessinger *et al.* eng. *Current Opinion in Structural Biology* **21**, 412 (2011).

[235] V. N. Uversky *et al. BMC Genomics* **10**, S7 (2009).

[236] V. N. Uversky and A. K. Dunker. eng. *Biochimica Et Biophysica Acta* **1804**, 1231 (2010).

[237] V. N. Uversky. eng. *Protein Science: A Publication of the Protein Society* **11**, 739 (2002).

[238] E. A. Cino *et al.* en. *PLOS ONE* **6**. Publisher: Public Library of Science, e27371 (2011).

[239] H. J. Dyson. eng. *Quarterly Reviews of Biophysics* **44**, 467 (2011).

[240] V. N. Uversky. en. *The International Journal of Biochemistry & Cell Biology* **43**, 1090 (2011).

[241] A. K. Dunker and Z. Obradovic. eng. *Nature Biotechnology* **19**, 805 (2001).

[242] F. Karush. *Journal of the American Chemical Society* **72**. Publisher: American Chemical Society, 2705 (1950).

[243] R. W. Kriwacki *et al.* eng. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 11504 (1996).

[244] M. Fuxreiter, P. Tompa, and I. Simon. en. *Bioinformatics* **23**. Publisher: Oxford Academic, 950 (2007).

[245] L. M. Iakoucheva *et al.* en. *Journal of Molecular Biology* **323**, 573 (2002).

[246] A. Campen *et al.* eng. *Protein and Peptide Letters* **15**, 956 (2008).

[247] S. J. Metallo. *Current opinion in chemical biology* **14**, 481 (2010).

[248] V. N. Uversky, C. J. Oldfield, and A. K. Dunker. eng. *Annual Review of Biophysics* **37**, 215 (2008).

[249] V. N. Uversky. *Frontiers in Aging Neuroscience* **7** (2015).

[250] V. N. Uversky. *Expert Review of Proteomics* **7**, 543 (2010).

[251] U. Midic *et al.* eng. *Protein and Peptide Letters* **16**, 1533 (2009).

[252] D. I. Hammoudeh *et al.* *Journal of the American Chemical Society* **131**. Publisher: American Chemical Society, 7390 (2009).

[253] C. Y *et al.* English. *Trends in Biotechnology* **24**, 435 (2006).

[254] J. Wang *et al.* *International Journal of Molecular Sciences* **12**, 3205 (2011).

[255] M. Zhu *et al.* eng. *The Journal of Chemical Physics* **139**, 035101 (2013).

[256] G. Tóth *et al.* en. *PLOS ONE* **9**. Publisher: Public Library of Science, e87133 (2014).

[257] S. Sinha *et al.* *Journal of the American Chemical Society* **133**. Publisher: American Chemical Society, 16958 (2011).

[258] S. Prabhudesai *et al.* eng. *Neurotherapeutics: The Journal of the American Society for Experimental NeuroTherapeutics* **9**, 464 (2012).

[259] M. Fokkens, T. Schrader, and F.-G. Klärner. *Journal of the American Chemical Society* **127**. Publisher: American Chemical Society, 14415 (2005).

[260] F. Jin *et al.* en. *PLOS Computational Biology* **9**. Publisher: Public Library of Science, e1003249 (2013).

[261] D. Kumar, N. Sharma, and R. Giri. *Cancer Informatics* **16** (2017).

[262] A. V. Follis *et al.* English. *Chemistry & Biology* **15**. Publisher: Elsevier, 1149 (2008).

[263] E. V. Prochownik. *Expert Review of Anticancer Therapy* **4**. Publisher: Taylor & Francis _: https://doi.org/10.1586/14737140.4.2.289, 289 (2004).

[264] D. M. Clausen *et al. The Journal of Pharmacology and Experimental Therapeutics* **335**, 715 (2010).

[265] S. Pelengaris, M. Khan, and G. Evan. en. *Nature Reviews Cancer* **2**. Number: 10 Publisher: Nature Publishing Group, 764 (2002).

[266] D. W. Felsher and J. M. Bishop. en. *Molecular Cell* **4**, 199 (1999).

[267] A. Siddiqui-Jain *et al.* eng. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 11593 (2002).

[268] P. Russo *et al.* eng. *The Journal of Pharmacology and Experimental Therapeutics* **304**, 37 (2003).

[269] X. Yin *et al.* en. *Oncogene* **22**. Number: 40 Publisher: Nature Publishing Group, 6151 (2003).

[270] M.-J. Huang *et al.* en. *Experimental Hematology* **34**, 1480 (2006).

[271] H. Mo and M. Henriksson. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 6344 (2006).

[272] C.-P. Lin *et al.* eng. *Anti-Cancer Drugs* **18**, 161 (2007).

[273] S. K. Nair and S. K. Burley. eng. *Cell* **112**, 193 (2003).

[274] M. Fladvad *et al.* eng. *Journal of Molecular Biology* **346**, 175 (2005).

[275] E. M. Blackwood and R. N. Eisenman. eng. *Science (New York, N.Y.)* **251**, 1211 (1991).

[276] R. Turner. en. *Nature Structural & Molecular Biology* **10**. Number: 3 Publisher: Nature Publishing Group, 157 (2003).

[277] C. Dang. en. *Cell* **149**, 22 (2012).

[278] J. L. Yap *et al.* en. *MedChemComm* **3**. Publisher: The Royal Society of Chemistry, 541 (2012).

[279] T. Berg *et al.* en. *Proceedings of the National Academy of Sciences* **99**. Publisher: National Academy of Sciences Section: Biological Sciences, 3830 (2002).

[280] J. Shi *et al.* eng. *Bioorganic & Medicinal Chemistry Letters* **19**, 6038 (2009).

[281] Y. Xu *et al.* eng. *Bioorganic & Medicinal Chemistry* **14**, 2660 (2006).

[282] A. Kiessling *et al.* eng. *ChemMedChem* **2**, 627 (2007).

[283] J. R. Hart *et al.* eng. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 12556 (2014).

[284] S. H. Choi *et al. ACS Chemical Biology* **12**. Publisher: American Chemical Society, 2715 (2017).

[285] A. Castell *et al.* en. *Scientific Reports* **8**. Number: 1 Publisher: Nature Publishing Group, 1 (2018).

[286] L. A. Carabet, P. S. Rennie, and A. Cherkasov. *International Journal of Molecular Sciences* **20** (2018).

[287] H. Wang *et al. Oncotarget* **4**, 936 (2013).

[288] I. Müller *et al. PLoS ONE* **9** (2014).

[289] G. T. Heller *et al.* en. *Journal of Molecular Biology* **429**, 2772 (2017).

[290] M. J and C. R. English. *Plos one* **7**, e41070 (2012).

[291] S. Ambadipudi and M. Zweckstetter. eng. *Expert Opinion on Drug Discovery* **11**, 65 (2016).

[292] V. S. Pande, K. Beauchamp, and G. R. Bowman. eng. *Methods (San Diego, Calif.)* **52**, 99 (2010).

[293] D. Song, R. Luo, and H.-F. Chen. *Journal of chemical information and modeling* **57**, 1166 (2017).

[294] B. Hess *et al.* en. *Journal of Computational Chemistry* **18**, 1463 (1997).

[295] G. Bussi, D. Donadio, and M. Parrinello. *The Journal of Chemical Physics* **126**. Publisher: American Institute of Physics, 014101 (2007).

[296] M. Parrinello and A. Rahman. *Journal of Applied Physics* **52**. Publisher: American Institute of Physics, 7182 (1981).

[297] U. Essmann *et al. The Journal of Chemical Physics* **103**. Publisher: American Institute of Physics, 8577 (1995).

[298] J. Huang *et al.* en. *Nature Methods* **14**. Number: 1 Publisher: Nature Publishing Group, 71 (2017).

[299] K. Vanommeslaeghe *et al. Journal of computational chemistry* **31**, 671 (2010).

[300] M. K. Scherer *et al. Journal of Chemical Theory and Computation* **11**. Publisher: American Chemical Society, 5525 (2015).

[301] I. Buch, T. Giorgino, and G. D. Fabritiis. en. *Proceedings of the National Academy of Sciences* **108**. Publisher: National Academy of Sciences Section: Biological Sciences, 10184 (2011).

[302] M. Bernetti *et al. Annual Review of Physical Chemistry* **70**. Publisher: Annual Reviews, 143 (2019).

[303] C. J. Brown *et al.* en. *Nature Reviews Cancer* **9**. Number: 12 Publisher: Nature Publishing Group, 862 (2009).

[304] B. Vogelstein, D. Lane, and A. J. Levine. en. *Nature* **408**. Number: 6810 Publisher: Nature Publishing Group, 307 (2000).

[305] K. H. Vousden and C. Prives. en. *Cell* **137**, 413 (2009).

[306] E. Kastenhuber and S. Lowe. *Cell* **170**, 1062 (2017).

[307] G. Sanz *et al.* en. *Journal of Molecular Cell Biology* **11**. Publisher: Oxford Academic, 586 (2019).

[308] C. J. Kemp *et al.* eng. *Cell* **74**, 813 (1993).

[309] A. Feki and I. Irminger-Finger. eng. *Critical Reviews in Oncology/Hematology* **52**, 103 (2004).

[310] G. Selivanova *et al.* eng. *Nature Medicine* **3**, 632 (1997).

[311] B. A. Foster *et al.* eng. *Science (New York, N.Y.)* **286**, 2507 (1999).

[312] A. C. Joerger, H. C. Ang, and A. R. Fersht. en. *Proceedings of the National Academy of Sciences* **103**. Publisher: National Academy of Sciences Section: Biological Sciences, 15056 (2006).

[313] V. J. N. Bykov *et al.* en. *Nature Medicine* **8**. Number: 3 Publisher: Nature Publishing Group, 282 (2002).

[314] K. H. Vousden and D. P. Lane. en. *Nature Reviews Molecular Cell Biology* **8**. Number: 4 Publisher: Nature Publishing Group, 275 (2007).

[315] Q. Li and G. Lozano. eng. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **19**, 34 (2013).

[316] S. Francoz *et al.* en. *Proceedings of the National Academy of Sciences* **103**. Publisher: National Academy of Sciences Section: Biological Sciences, 3232 (2006).

[317] M. Wade, Y.-C. Li, and G. M. Wahl. *Nature reviews. Cancer* **13**, 83 (2013).

[318] P. H. Kussie *et al.* eng. *Science (New York, N.Y.)* **274**, 948 (1996).

[319] J. D. Oliner *et al.* en. *Nature* **362**. Number: 6423 Publisher: Nature Publishing Group, 857 (1993).

[320] L. T. Vassilev *et al.* eng. *Science (New York, N.Y.)* **303**, 844 (2004).

[321] B. Vu *et al.* eng. *ACS medicinal chemistry letters* **4**, 466 (2013).

[322] Q. Ding *et al.* *Journal of Medicinal Chemistry* **56**. Publisher: American Chemical Society, 5979 (2013).

[323] K. Ding *et al.* *Journal of Medicinal Chemistry* **49**. Publisher: American Chemical Society, 3432 (2006).

[324] Y. Zhao *et al.* eng. *Journal of Medicinal Chemistry* **56**, 5553 (2013).

[325] S. Wang *et al.* eng. *Cancer Research* **74**, 5855 (2014).

[326] B. L. Grasberger *et al.* eng. *Journal of Medicinal Chemistry* **48**, 909 (2005).

[327] D. Sun *et al.* *Journal of Medicinal Chemistry* **57**. Publisher: American Chemical Society, 1454 (2014).

[328] Y. Rew *et al.* eng. *Journal of Medicinal Chemistry* **55**, 4936 (2012).

[329] M. A. McCoy *et al.* en. *Proceedings of the National Academy of Sciences* **100**. Publisher: National Academy of Sciences Section: Biological Sciences, 1645 (2003).

[330] S. A. Showalter *et al. Journal of the American Chemical Society* **130**. Publisher: American Chemical Society, 6472 (2008).

[331] J. A. Bueren-Calabuig and J. Michel. *PLoS Computational Biology* **11** (2015).

[332] K. Michelsen *et al. Journal of the American Chemical Society* **134**. Publisher: American Chemical Society, 17059 (2012).