



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

Scienze Biotecnologiche, Biocomputazionali,
Farmaceutiche e Farmacologiche

Ciclo XXXVIII

Settore Concorsuale: 03/D1 – Chimica e Tecnologie Farmaceutiche, Tossicologiche e Nutraceutico-Alimentari

Settore Scientifico Disciplinare: CHIM/08 – Chimica Farmaceutica

Bidirectional path-based nonequilibrium simulations for binding free energy estimation

Presentata da: Eleonora Serra

Coordinatore Dottorato

Prof.ssa Maria Laura Bolognesi

Supervisore

Prof. Andrea Cavalli

Co-supervisore

Dr. Sergio Decherchi

Esame Finale – Anno 2026

*“Ho sceso, dandoti il braccio, almeno un milione di scale,
e ora che non ci sei è il vuoto ad ogni gradino.”*

– ***Eugenio Montale***, *Xenia II*, n°5

**Alle persone che sono state, sono e saranno sempre
parte di questo mio strano viaggio chiamato vita.**

Contents

Abbreviations and Symbols	9
0.1 Abbreviations	9
0.2 Symbols and Constants	10
Abstract	12
Preface	13
1 Introduction	16
1.1 Thermodynamics & Binding Free Energy	22
1.2 Kinetics & Binding Kinetics	28
1.2.1 Kinetics in Drug Discovery	32
2 Theoretical background	34
2.1 Statistical mechanics	35
2.1.1 Statistical ensembles	38
2.1.2 Molecular systems	42
2.2 Molecular Mechanics	43
2.3 Molecular Dynamics	45
2.3.1 Integration Algorithm	48
2.3.2 Interactions treatment and Force Fields	52
2.3.3 System preparation	61
2.3.4 Thermostats and Barostats	63
2.3.5 Periodic Boundary Conditions (PBC)	66

2.3.6	Water models	67
2.3.7	MD output and conclusion	74
2.4	Rare events & Enhanced sampling	75
2.4.1	Order parameter: coupling parameter & CV	76
2.4.2	Enhanced sampling methods	79
2.5	Binding Free Energy Calculations	84
2.5.1	Sampling Protocol	87
2.5.2	Free Energy estimators	90
3	On free energy calculations in drug discovery [P5]	97
3.1	Introduction	97
3.2	Alchemical Transformations in Drug Discovery	99
3.3	Path Based methods in Drug Discovery	102
3.3.1	Umbrella sampling	103
3.3.2	Metadynamics	105
3.3.3	Steered Molecular Dynamics	106
3.4	Challenges and perspective	109
4	Bidirectional Path-based nonequilibrium simulations for binding free energy [P1]	113
4.1	Aim and Objectives	113
4.2	Introduction	113
4.3	Methods	114
4.3.1	Path collective variables and Path definition	115
4.3.2	Nonequilibrium Simulations and Free Energy Estimation . .	117
4.3.3	Overall Computational pipeline	118
4.4	Results and Discussion	122
4.4.1	CB8-G8 System	122
4.4.2	β -Trypsin-Benzamidine System	123
4.4.3	Abl-Gleevec System	133
4.5	Conclusion	136
5	Minimizing dissipation in nonequilibrium binding free energy sim-	

ulations [P2]	138
5.1 Aim and Objectives	138
5.2 Introduction	139
5.2.1 Water Models	141
5.2.2 Path collective variables	142
5.3 Methods	143
5.3.1 Selected systems	143
5.4 Results and Discussion	145
5.4.1 Dissipation due to water model	146
5.5 Dissipation due to PCVs	160
5.5.1 Reference path definition	161
5.5.2 Orthogonal deviation from the path	168
5.6 Conclusion	170
6 Expanding the applicability of nonequilibrium binding free energy	
estimation: from protein to RNA [P3]	172
6.1 Aim and Objectives	172
6.2 Introduction	173
6.2.1 Systems	173
6.3 Methods	176
6.3.1 ABMD and PCVs	176
6.3.2 SMD and PMF estimation	177
6.3.3 Standard Binding Free Energy Estimation	178
6.3.4 Well-Tempered MetaDynamics simulations	178
6.3.5 Setup of the systems	179
6.4 Results and Discussion	182
6.4.1 Abl-Gleevec: ligand flexibility and large conformational re-	
arrangements	183
6.4.2 Riboswitch-ligand: Receptor Flexibility and Conformational	
Rearrangements	192
6.5 Conclusions	210
7 NanoShaperWeb: molecular surface and pockets detection made	

visual [P4]	212
7.1 Aim and Objectives	212
7.2 Introduction	213
7.3 Implementation	214
7.3.1 NanoShaper and Descriptors	215
7.3.2 Web Server	217
7.3.3 Browser App	218
7.4 Web server interface	219
7.4.1 Input, Settings, Viewer and Pocket Table	220
7.4.2 Mail Notification and Results Download	224
7.5 Conclusion	225
8 Absolute residence times estimation from scaled molecular dynamics simulations: application to Mcl-1 inhibitors	226
8.1 Aim and Objectives	226
8.2 Introduction	227
8.3 Methods	228
8.3.1 From Scaled to Relative Unbinding Times	228
8.3.2 From Scaled to Absolute Unbinding Times	229
8.3.3 Relationship between scaled and physical RT	229
8.4 Results and Discussion	232
8.4.1 Calibration of the Scaling Factor	232
8.4.2 Relative and Absolute Unbinding Time Predictions	233
8.5 Conclusion	234
9 Supporting Information	235
9.1 Minimizing dissipation is nonequilibrium binding free energy simulations	235
9.1.1 Well-Tempered MetaDynamics	235
9.2 Expand the applicability of nonequilibrium binding free energy estimation: from protein to RNA	239
9.2.1 Additional results for RNA-ligand systems	239
9.2.2 Debye-Hückel interaction energy profile	240

9.2.3 Analysis of PCV and Path definition	240
10 Conclusions	242
Publications	247
Bibliography	248
Ringraziamenti	276

Abbreviations and Symbols

The most important and frequently used abbreviations and symbols in this thesis are listed below, along with their explanations.

0.1 Abbreviations

- NMR: nuclear magnetic resonance
- MD: molecular dynamics
- CADD: computer-aided drug design
- RT : residence time
- QM: Quantum Mechanics
- MM: Molecular Mechanics
- NVE : Microcanonical ensemble
- NVT : Canonical ensemble
- NPT : Isothermal isobaric ensemble
- $T\mu V$: Grand canonical ensemble
- FF: force field
- PES: Potential energy surface
- FES: Free energy surface
- PMF: Potential of Mean Force

- PBC: Periodic Boundary conditions
- LJ: Lennard-Jones potential
- TIPNP: Transferable Intermolecular Potential N-Point model
- OPC: Optimal Point Charge model
- CV: Collective Variable
- PCV: Path Collective Variable
- RMSD: Root-mean-square deviation
- MSD: Mean-square deviation
- US: Umbrella Sampling
- MetaD, WT-MetaD: Metadynamics, Well-Tempered MetaD
- SMD: Steered Molecular Dynamics
- ABMD: Adiabatic Bias Molecular Dynamics
- TI: Thermodynamic Integration
- FEP: Free Energy Perturbation
- BAR: Bennett Acceptance Ratio method
- JE: Jarzynski equality
- CFT: Crooks Fluctuation Theorem
- AM1-BCC: Austin Model 1 semi-empirical QM calculations with Bond Charge Corrections
- RESP: Restrained Electrostatic Potential
- PME: Particle Mesh Ewald

0.2 Symbols and Constants

- k_B : Boltzmann constant ($k_B = 1.380649 \cdot 10^{-23} JK^{-1}$)

- $\beta = \frac{1}{k_B T}$: inverse thermal energy
- k_{on} and k_{off} : association and dissociation rate constants
- ΔG_b° : Gibbs standard binding free energy
- K_d and K_a : dissociation and association constants
- C° : standard-state concentration
- G, F : Gibbs free energy, Helmholtz free energy
- H, S : Enthalpy, Entropy
- U, K : Potential Energy, Kinetic Energy
- T, P, V : Temperature, Pressure, Volume
- R : universal gas constant
- $\Delta G_b, \Delta H_b, \Delta S_b$: binding free energy, binding enthalpy, binding entropy.
- $\Delta S_{conf}, \Delta S_{solv}, \Delta S_{roto/trasl}$: conformational, solvation, roto/traslational entropy
- E_a : activation energy
- A : Arrhenius proportionality constant
- κ = Transmission coefficient
- $\Delta G^\ddagger$: activation Gibbs free energy
- δ : Dirac function
- h : Planck constant ($h = 6.626\,070\,15 \times 10^{-34} \text{ J} \cdot \text{Hz}^{-1}$)
- $\hat{H}$: Hamiltonian
- Q_{NVT} : canonical partition function
- Z_{NVT} configurational integral
- W_J : Jarzynski work
- W_{diss} : Dissipated Jarzynski work

Abstract

Drug discovery is an inherently complex, time-consuming, and costly process, often requiring over a decade of research and investments amounting to billions of dollars. A key quantity that underpins the success of early-stage drug design is the binding free energy, which quantifies the strength of interactions between a protein and a potential ligand. Accurate estimation of binding free energy is essential for identifying promising drug candidates and prioritizing them for experimental validation. This thesis first gives a comprehensive examination of the theoretical principles that govern simulation for binding free energy estimation, and then introduces novel protocols. We give a detailed review of the statistical mechanics framework for binding free energy calculations and critically evaluate key free energy estimators by analyzing their underlying assumptions, practical advantages, and limitations. In addition to consolidating theoretical knowledge, this work contributes practical insights by assessing the applicability and the performance of some of these estimators, particularly nonequilibrium estimators, across representative molecular systems. The findings support and clarify the use of nonequilibrium approaches under diverse conditions and in increasingly complex systems, while also providing guidance for best practices in real-world drug discovery pipelines. Taken together, these contributions strengthen the theoretical foundations of protein-ligand binding free energy estimation and give valuable tools and guidelines to improve the efficiency of computational drug discovery efforts.

Preface

Thermodynamics and kinetics play crucial roles in determining the dynamic behaviors and function of biological molecules within the human body. Therefore, differences in thermodynamic binding free energy and the kinetic properties of the biological binding and unbinding processes (i.e., association between a ligand and a macromolecular target) are among the most critical quantities for the optimization of new drug candidates.

The drug discovery and development process, responsible for bringing new therapeutics to market, is both lengthy and resource-intensive, requiring significant financial investments from pharmaceutical companies, governments, and other funding bodies. In recent decades, *in-silico* methods have become essential tools in this field. Computational simulations, in particular, play a crucial role in analyzing and optimizing potential drug candidates. They enable the estimation of binding affinities to target molecules and kinetic rates, helping to streamline drug development while reducing costs and risks.

Among all biological molecular partners, proteins play a crucial role, with nearly all significant human diseases being linked to protein dysfunction. The binding of small molecules to proteins is a key process in cellular metabolism, and it is also exploited in pharmacology to treat a wide range of diseases. Typically, drug development strategies focus on improving ligand affinity to a specific target protein, guided by the structure and interactions observed in ligand-protein complexes. However, binding kinetics and ligand residence time are also crucial factors in determining a drug's physiological response and efficacy. Indeed, a clinically effective

drug response is based on the ability of the therapeutic molecule to bind to its biological target and the subsequent translation of this interaction into a selective physiological effect. Consequently, the critical prerequisite for drug action is its binding to the biological target as well as the duration of this binding.

The three-dimensional structure of biomolecules and their conformational landscape provides invaluable insights into biological processes, facilitating the design of new drug candidates. As a result, tremendous efforts have been invested in properly determining the three-dimensional structures of key biomolecules. Experimental techniques like X-ray crystallography and nuclear magnetic resonance (NMR) are widely employed by biologists to resolve and visualize them. However, due to experimental limitations, computational simulations have become essential complementary methods to study biological processes in detail, including conformational changes, protein folding and ligand binding. Particularly, atomic-level molecular dynamics (MD) simulations are the method of choice for protein simulations. They enable researchers to observe the movements of individual atoms within biological molecules, providing atomic-resolution information that is often unattainable through traditional experimental approaches.

In the context of protein-ligand binding, the thermodynamics of the process is generally quantified using binding free energies or equilibrium dissociation constants, while its kinetics is characterized by association and dissociation rate constants. Experimental biophysical techniques such as isothermal titration calorimetry (ITC), surface plasmon resonance (SPR), biolayer interferometry (BLI), fluorescence spectroscopy, and grating-coupled interferometry (GCI) are commonly employed to quantitatively characterize these parameters. On the other hand, computational approaches to quantitatively assess these quantities are still facing significant challenges, mainly due to the complexity of the analyzed systems and the considerable computational resources required to simulate the protein-ligand association process. Therefore, the existing computational methods often face limitations in practical application to real pharmaceutical systems, particularly within the time constraints of drug discovery processes.

In this context, this PhD research focused on exploring and estimating the ther-

modynamics of protein-ligand complexes through *in-silico* approaches, alongside a further side project investigating binding kinetics and other contributions. Given the potential of nonequilibrium approaches, the research focused on developing a novel computational pipeline for calculating binding free energies in nonequilibrium scenarios (Chapter 4 [P1]). While nonequilibrium methods offer inherent parallelization and computational advantages, they also present challenges related to dissipation and convergence issues. Thus, we deeply investigated the underlying causes of dissipation in our nonequilibrium protocol, defining practical guidelines to improve the accuracy of binding free energy estimations (Chapter 5 [P2]). In parallel, the work concentrated on extending the applicability of the methodology to very complex systems, ranging from challenging proteins to inherently flexible RNA complexes (Chapter 6 [P3]). Chapter 8 and Chapter 7 focus on more side projects: Chapter 8 is dedicated to the kinetic study of Mcl-1 protein and its inhibitors, while Chapter 7 presents the development of NanoShaperWeb [P4], a freely accessible online tool based on the NanoShaper code for computing molecular surfaces and pockets. These chapters, which describe the carried out research, are prefaced by Chapter 1 which introduces fundamental thermodynamic and kinetic concepts in the framework of drug discovery. Chapter 2 details the theoretical background of simulation approaches used to characterize protein-ligand interactions and to predict thermodynamic and kinetic properties. Finally, this introductory set of Chapters, is closed by Chapter 3 which provides an overview of our comprehensive account on binding free energy estimation approaches in drug discovery, with a specific focus on the protein-ligand binding process [P5]. It gives both an analytical and critical examination of various free energy methods, highlighting their applications and relevance within the context of our laboratory research. Building on the key conclusions presented in the account and the developed methods, the thesis closes with a discussion of the current open challenges in the field. This is provided in Chapter 10.

Chapter 1

Introduction

The rational drug discovery process, responsible for bringing new drugs to market, follows five key stages [1]:

1. The pre-discovery phase, where basic research is conducted to understand disease mechanisms and identify potential biological targets, such as proteins. It involves the target selection and validation.
2. The drug discovery step which explores exogenous molecules (primarily small molecules or biologics) that could potentially treat the disease. During this phase, the hit-to-lead and lead optimization processes are crucial, as they refine initial hit compounds into more promising candidates suitable for clinical development.
3. The preclinical development, where scientists study the potential drug's mechanism assessing toxicity, validating its effectiveness (*in vitro* and *in vivo* models), and exploring formulation option.
4. The clinical trials, where the drug is tested in humans with pharmacovigilance studies.
5. The final step including regulatory review, approval, and post-market surveillance.

As one might expect, the development of new drugs demands substantial financial

investments from both pharmaceutical companies and governments [2]. For every new drug approved for human use, approximately 5,000 to 10,000 chemical compounds undergo rigorous chemical and biological screening. Of these, around 250 will progress to preclinical testing, and only 5 will eventually enter clinical trials. Thus, bringing a new therapeutic to market is a long-lasting process, typically taking 10 to 15 years, with estimated costs ranging from 1.5 to 2.0 billion US dollars [3, 4, 1]. A significant part of research efforts and financial investment is dedicated to fully understand the disease mechanisms at the molecular level, as well as the potential effects of exogenous drug candidates. However, despite the technological advancements and a deeper understanding of biological systems, accurately predicting how a living organism may respond to a potential drug remains a great challenge. Improving these predictive capabilities is therefore essential, as it can reduce the time and cost associated with rational drug discovery process [5].

In human body, biological macro-molecules and their transient aggregation (with each other or with small molecules) govern and mediate all relevant biochemical processes. During this aggregation process, molecular recognition is considered the basic prerequisite for two molecules to interact [6]. Molecular recognition is controlled by specific non-covalent interactions and depends on both specificity and affinity. Specificity allows a molecule to selectively bind one partner, while affinity measures the binding strength [7].

Among all biological interacting partners, proteins play crucial roles in a wide numbers of cellular processes, including structural, mechanical, and cell signaling functions [8, 7]. Notably, proteins perform their biological functions through physical interactions with other molecules, such as other proteins, peptides or small molecule ligands [9, 8, 7]. Approximately 95% of the total receptors in the human body are proteins, where receptors are defined as specific macromolecules to which a ligand binds by altering their biochemical activity [9, 8]. Given their central role in the human body, many diseases are directly associated with dysfunctional proteins [4]. However, a large proportion of disease-related human proteins are considered undruggable by current therapeutic approaches. A significant class of these undruggable targets includes proteins involved in intracellular protein-protein

interactions, which represent an open challenge for rational drug development [10]. For this reason, computer-aided drug design (CADD) has become an indispensable tool for facilitating and accelerate the complex process of designing and optimizing new drug candidates and novel therapeutic approaches [5, 11].

The interaction of a protein with small molecules is a crucial aspect of cellular metabolism, and it is extensively utilized in pharmacology for the treatment of various diseases [12, 13, 9]. In the framework of rational drug discovery, a small molecule that binds to a biological target, with reasonable specificity and affinity, is referred to as a ligand. The subsequent biological response is a result of chemical interactions between the ligand and the protein’s binding site [9, 7]. From this point onward, the term ligand will specifically refer to any molecule capable of binding to a protein [7]. To describe the physicochemical mechanisms of the protein-ligand binding process, three different models exists [7]:

- lock-and-key model [14]: in this model introduced by Emil Fischer, both the protein and the ligand are rigid entities, requiring their binding surfaces to align perfectly. As a result, only a ligand with the correct dimensions (the key) can properly fit into the protein’s binding site (the lock). However, this simplified model fails to explain experimental evidence showing that a protein can bind to its ligand even when their initial shapes are not a perfect match.
- Induced-fit model [15]: in this model, the protein’s binding site is considered flexible, and the ligand induces a conformational change upon its binding. Since the induced-fit model focuses solely on the flexibility of the ligand-binding site, it is appropriate mostly for proteins that undergo minor conformational changes after ligand binding.
- Conformational selection model [16]: this more recent model recognizes that most proteins are inherently dynamic, existing as an ensemble of conformational states in equilibrium. This model suggests that a ligand binds selectively to the most suitable protein conformation, shifting the equilibrium toward this state. Even in the unbound state, the protein can sample conformations similar to those adopted when it is bound to a ligand.

Nowadays, the original and rigid Fischer’s lock-and-key binding paradigm has been surpassed by the other two models; i.e. the induced-fit and conformational selection mechanisms. It is widely recognized that both mechanisms contribute to molecular recognition, offering a more dynamic and exhaustive understanding of biomolecular interactions [16]. A comprehensive discussion about the physicochemical mechanisms of the protein-ligand binding process can be found in references [17, 7, 18].

In this context, the interaction between a protein and its ligand is known to depend on their random coupling [4]. To simplify the discussion, we focus on a fundamental two-state mechanism characterized by a single elementary step, ignoring any intermediate states. During this protein-ligand binding process, the protein and ligand exist in equilibrium with the protein-ligand complex, as illustrated in Figure 1.1 and represented below:



Here, $[P]$ is the free concentration of the receptor protein, $[L]$ is the free concentration of the ligand, while $[PL]$ is the concentration of the binary protein-ligand complex. Square brackets indicate the concentration of any molecular species, usually expressed in mole/liter (M). Moreover, k_{on} and k_{off} are the rate constants for the association and dissociation process of the binary complex. The units of k_{on} and k_{off} are $M^{-1}s$ and s^{-1} , respectively.

Equilibrium is achieved when the rate of formation of the protein-ligand complex equals the rate of its dissociation, meaning that the forward binding reaction is balanced by the reverse unbinding reaction:

$$k_{\text{on}}[P][L] = k_{\text{off}}[PL] \quad (1.2)$$

Starting from this equilibrium, the standard binding free energy $\Delta G_{\text{binding}}^{\circ}$ (i.e., the free energy difference between the bound and unbound states) can be evaluated:

$$\Delta G_{\text{binding}}^{\circ} = G_{PL}^{\circ} - (G_P^{\circ} + G_L^{\circ}) \quad (1.3)$$

Notably, standard $\Delta G_{\text{binding}}^{\circ}$ refers to the free energy difference measured under

standard conditions, which are 1 *atm* pressure, temperature of 298 *K*, and the effective reactant concentrations of 1 *M*. $\Delta G_{binding}^\circ$ is a direct measurement of the binding affinity and, at equilibrium, it is related to the dissociation constant K_d (unit of *M*):

$$\Delta G_{binding}^\circ = k_B T \ln \frac{K_d}{C^\circ} \quad (1.4)$$

where C° is the standard-state concentration (1 *M*), k_B is the Boltzmann constant, and K_d represents the inverse of the receptor's affinity for the ligand at equilibrium. Therefore, when K_d is large, it indicates that the ligand does not bind easily to the receptor; conversely, if K_d is small, the ligand binds readily to its target. K_d quantifies the ligand concentration at equilibrium where the protein is equally likely to be in its bound or unbound state:

$$K_d = \frac{[P][L]}{[PL]} \quad (1.5)$$

The inverse of K_d is the association constant K_a (unit of M^{-1}), which characterizes the binding process and it is the ratio between the complex concentration and the concentrations of the ligand and protein free forms.

Thermodynamics and kinetics of binding are both associated with the energetic profiles of the molecular interactions under investigation. Thermodynamics defines the energetic driving force behind the binding process, determining the stability of the complex. On the other hand, kinetics describes the rates of transitions between energy minima, governing how quickly a ligand associates with or dissociates from its target. Therefore, thermodynamic and kinetic constants are related with different phenomena. From a thermodynamic perspective, K_d is directly linked to the stable intermolecular interactions formed between the ligand, the protein, and the solvent. Conversely, the kinetic rate constants are influenced by transient interactions that occur during the binding process. These kinetic rate constants correspond to the highest free energy barrier, known as the transition state, that separates the bound and unbound states. Interestingly, binding free energy is also related to the free energy of this transition state. Thus, the thermodynamic and

kinetic constants of binding are interconnected through the following relationship:

$$K_d = \frac{[P][L]}{[PL]} = \frac{k_{off}}{k_{on}} \quad (1.6)$$

Accordingly, a fast binding rate combined with a slow dissociation rate will result in a low dissociation constant and, hence, a high binding affinity.

In conclusion, a deep understanding of ligand action at the molecular level is essential for guiding the rational design of new therapeutics. This process is informed by insights into both the kinetics and thermodynamics of the ligand binding process [13, 19, 20, 4]. In the following sections, fundamental thermodynamic principles and their relevance to drug discovery (particularly in the context of protein-ligand binding) will be introduced, followed by a discussion of the critical role of kinetics in optimizing the drug discovery process.

Protein-ligand binding process

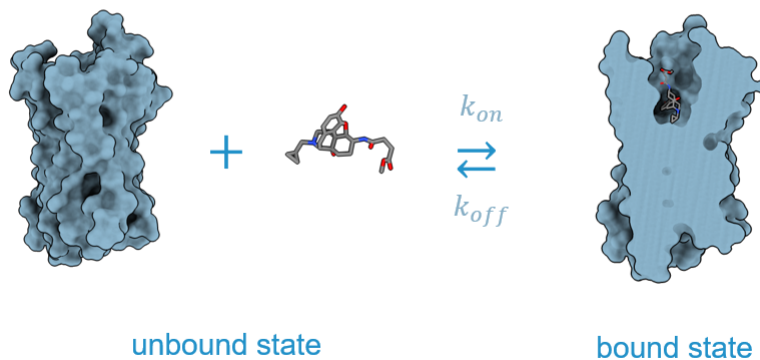


Figure 1.1: Schematic representation of the protein-ligand binding event. Starting from the protein and ligand in their free forms, the two partners interact and bind, forming a binary complex that represents the ligand’s bound state. In this depiction, the protein is shown as a molecular surface to emphasize the binding pocket, while the ligand is illustrated in a licorice representation.

1.1 Thermodynamics & Binding Free Energy

The basic thermodynamic system made up of a solvent and a solute serves as a simplified model for a more complex system that includes the protein, the ligand, and the solvent molecules. In this system, molecules interact through complex interactions and heat exchange. The relationships between these molecules and the energy exchanges are governed by the laws of thermodynamics [21].

In the framework of thermodynamics, whether a process occurs spontaneously is determined by its free energy. Indeed, systems tend to evolve toward their lowest free energy state in the absence of any external energy source. Free energy is a measure of the probability of finding a system in a given state. For this reason, free energies are crucial thermodynamic quantities to determine binding affinities, influencing whether binding occurs spontaneously or not [4].

Generally, free energy is the physical quantity minimized at equilibrium, and it represents the maximum energy obtainable from a spontaneous process at constant T . In other words, it is the minimum energy needed to drive a process that goes against its natural direction. Free energy is the energy available from a process that occurs slowly enough to remain in equilibrium with its surroundings, i.e., a reversible process. If the process is forced to take place faster, the work done on the system will exceed the change in free energy due to friction, and this extra energy will be lost in the form of heat. In this case, the dissipated work leads to the irreversibility of the process [22]. Two different flavors of free energies can be defined, namely the Helmholtz and the Gibbs free energies [23].

The Gibbs free energy is defined as:

$$G = H - TS \tag{1.7}$$

with T the temperature, S the entropy, and H the enthalpy of the system. H is defined as $H = U + PV$, where P is the pressure and V the system's volume. Gibbs free energy is defined as the thermodynamic potential that measures a system's capacity to perform maximum or reversible work under constant conditions of T

and P (isothermal and isobaric conditions).

On the other hand, the Helmholtz free energy is defined as:

$$F = U - TS \quad (1.8)$$

with U the potential energy of the system. The potential energy quantifies the stored energy of a system as a result of its position in a force field. Differently from Gibbs free energy, Helmholtz free energy is applicable at constant V and T .

According to the previous formulas, free energies, hence affinities, are influenced by changes in interatomic interactions (through the internal or potential energy term H/U) and entropic fluctuations (as represented by the entropy term S). Notably, variations in solvent entropy can have a significant or even dominant impact on free energy. The PV term distinguishes Gibbs free energy from Helmholtz free energy. It quantifies the volume changes in Gibbs free energy, which is applicable at constant pressure, unlike Helmholtz free energy, which is relevant at constant volume. Under physiological conditions and common experimental setups, isobaric conditions are generally maintained; therefore, the Gibbs free energy gradient characterizes the driving force for the system's evolution.

Similar to any spontaneous process, protein-ligand binding occurs only when the change in Gibbs free energy ($\Delta G_{binding}$) of the system is negative at equilibrium. The magnitude of the negative $\Delta G_{binding}$ indicates the extent of protein-ligand association, thereby determining the ligand's binding affinity for a specific receptor and the stability of the resulting protein-ligand complex. It is important to note that free energy is a state function, which means that the differences in binding free energy depend solely on the initial and final thermodynamic states, regardless of the pathway that connects them, this holds at variance of kinetics which is path-dependent.

In addition to the definition of $\Delta G_{binding}^\circ$ as function of the dissociation constant K_d in Eq. 1.4, $\Delta G_{binding}$ can also be defined at any point during an association binding event (i.e., not just under standard-state conditions):

$$\Delta G_{binding} = \Delta G_{binding}^\circ + k_B T \ln c \quad (1.9)$$

where $\Delta G_{binding}$ represents the free energy change at any moment during the binding process, reflecting real-time variations in binding affinity as conditions evolve. In equation 1.9, c represents the reaction quotient which quantifies the ratio between the concentration of the protein-ligand complex and the product of the concentrations of the reactants at any moment (i.e., not mandatory at equilibrium). Furthermore, $\Delta G_{binding}$ is not only governed by structural factors (as in the equation 1.4), but also involves dynamic aspects.

$\Delta G_{binding}$ can be decomposed into its enthalpic and entropic contributions, according to the fundamental equation:

$$\Delta G_{binding} = \Delta H_{binding} - T\Delta S_{binding} \quad (1.10)$$

where $\Delta H_{binding}$ and $\Delta S_{binding}$ represent changes in enthalpy and entropy upon ligand binding. This equation emphasizes that the recognition event between two molecules is associated with changes of both the structure and dynamics of each interacting partner [24]. For simplicity, we will abbreviate the subscript *binding* as *b* in all equations from this point forward.

The first term in equation 1.10 is the enthalpic contribution to ΔG_b . Enthalpy represents the total energy of a thermodynamic system, encompassing both the internal energy of the solute and solvent, as well as the energy needed to create space for the system within the solvent. This *cavitation* term is the product between the volume and pressure. ΔH_b can be positive or negative: when ΔH_b is positive, it defines an endothermic process, while when ΔH_b is negative it defines an exothermic process. Considering a ligand binding process, an endothermic process means that energetically favorable non-covalent interactions are disrupted. On the other side, during an exothermic process, energetically favorable non-covalent interactions between atoms are formed. The enthalpic contribution to the binding free energy represents the system's energy change upon the ligand binding event (i.e., the intrinsic contribution of the molecular bonds formation), reflecting the specificity and strength of the non-covalent interactions between all partners. Indeed, ΔH_b is a global property of the system, including contributions from both the solute and the solvent. Ligand-receptor association involves the formation and dissociation

of numerous bonds involving both reagents and solvent molecules. In particular, this energy variation is linked to the formation of favorable non-covalent interactions at the bonding interface. These mainly include ionic, halogen and hydrogen bonds, electrostatic Coulomb and van der Waals interactions and other polar and non-polar interactions. However, the enthalpy change is also related to the loss of hydrogen bonds and van der Waals interactions between ligand and solvent, as well as between the protein and solvent. Additionally, it reflects the reorganization of solvent molecules near the surface of the newly formed complex. Overall, ΔH_b arises from the formation and disruption of numerous individual interactions, each of which can contribute either positively or negatively to the net enthalpy change. The net enthalpy change is a result of the cumulative effects of all these contributions.

The second term of equation 1.10 is the entropy contribution to ΔG_b . The straightforward explanation of this entropic part is that it reflects the system's overall dynamics, capturing how molecular motions and degrees of freedom change throughout the process. Entropy quantifies how the heat changes are distributed throughout the system. According to the second law of thermodynamics, heat naturally flows from higher T areas to lower T ones. This heat flow leads to a decrease in the system's initial order, which makes entropy a measure of the system's disorder or randomness. As the change in enthalpy, ΔS_b is a state function, and it is a global thermodynamic property. ΔS_b can be positive or negative: a positive change in entropy signifies an increase in the system's disorder, while a negative value of ΔS_b indicates a reduction in the system's degrees of freedom. Considering a ligand binding process, the changes in the binding entropy can be decomposed as a sum of three entropic contributions: conformational entropy ΔS_{conf} , solvation entropy ΔS_{solv} , and roto/translational entropy $\Delta S_{roto/transl}$. Binding reactions are likely to occur when the entropic penalties from $\Delta S_{roto/transl}$ are offset by significant positive contributions from ΔS_{conf} and ΔS_{solv} . These three entropic contributions are related to different process:

- ΔS_{conf} : the changes in conformational entropy reflect variations in the flexibility and movement of both the protein and ligand as they form a complex.

These fluctuations can either support or oppose the overall binding entropy, depending on whether the degree of freedom increases or decreases compared to the unbound states of the protein and ligand. When a drug-like molecule binds to a protein, it often becomes less mobile, leading to a reduction in ΔS_{conf} , which works against the forces promoting the binding interaction. However, ΔS_{conf} may be also favorable, and in some cases it may reduce the entropic cost of binding [25, 24, 26]

- ΔS_{solv} : this term quantifies the solvent entropy change. It mainly arises from the reduction of the exposed surface area, which leads to the displacement of solvent molecules during the binding event. This displacement usually contributes favorably to the binding entropy, as the release of solvent results in a significant positive entropy gain, thereby facilitating the binding interaction.
- $\Delta S_{roto/transl}$: this contribution reflects the loss of motion resulting from changes in the translational and rotational degrees of freedom of the interacting partners during binding. The complex formation decreases the number of free particles in solution, which negatively impacts the binding entropy. Molecular association leads in loss of certain degrees of freedom, particularly those associated with free translations and rotations.

Ultimately, the variations in enthalpy ΔH_b and entropy ΔS_b dictate the sign and magnitude of the binding free energy difference. The protein-ligand binding event occurs spontaneously only when the change in the system's free energy is negative, with its magnitude determining the stability of the resulting binary complex. Therefore, the two thermodynamics quantities ΔH_b and ΔS_b are the driving factors for protein-ligand binding, and their contributions are closely related. ΔG_b may have different thermodynamic signatures, varying from enthalpy- to entropy-driven.

Lastly, free energy can also be interpreted from a statistical perspective, where it plays a fundamental role in statistical mechanics, as further elaborated in Section 2.1. Within this framework, the probability of a system being in state χ is directly proportional to its free energy $G(\chi)$:

$$\rho(\chi) \propto e^{-\frac{G(\chi)}{k_B T}} \quad (1.11)$$

where the term $e^{-\frac{G(\chi)}{k_B T}}$ is the Boltzmann factor, while k_B is the Boltzmann constant. The latter is a fundamental physical constant that links microscopic particle behavior to macroscopic thermodynamic properties. It defines the amount of thermal energy per particle per degree of temperature and serves as a crucial bridge between statistical mechanics and thermodynamics. It is defined as:

$$k_B = \frac{R}{N_A} \quad (1.12)$$

where N_A is the Avogadro constant, while R is the universal gas constant. The precise value of the Boltzmann constant is $k_B = 1.38064910^{-23} JK^{-1}$.

In most practical cases, what one seeks to determine is not the absolute free energy value $G(\chi)$ but rather the free energy difference (ΔG) between two specific states, i.e. state A and state B . In the context of protein-ligand binding, usually these states correspond to the bound (i.e., state A) and the unbound states (i.e., state B). The previous relation becomes:

$$\frac{\rho(A)}{\rho(B)} = e^{-\frac{[G(A)-G(B)]}{k_B T}} = e^{-\frac{\Delta G_{AB}}{k_B T}} \quad (1.13)$$

leading to:

$$\Delta G_{AB} = G_B - G_A = -\frac{1}{\beta} \ln \frac{\rho(B)}{\rho(A)} \quad (1.14)$$

where $\beta = \frac{1}{k_B T}$ is the inverse thermal energy.

Equation 1.14 provides the most straightforward formula for estimating free energies from computer simulations. In principle, the free energy difference between two states can be theoretically estimated by directly evaluating the populations of states A and B using this relationship. With the same spirit, at constant T and V , ΔF between these two states is defined as:

$$\Delta F_{AB} = F_B - F_A = -\frac{1}{\beta} \ln \frac{\rho(B)}{\rho(A)} \quad (1.15)$$

A more detailed discussion on binding free energy calculations from simulations is provided in Chapter 2.5.

Finally, although a negative free energy difference and strong affinity for a target are essential for potential drug candidate, thermodynamics alone is insufficient to fully characterize protein-ligand binding under *in vivo* conditions. While optimizing the binding affinity, the energy of the fundamental state associated with the binary complex can unpredictably influence the binding kinetics. Therefore, both thermodynamic and kinetic aspects of molecular interactions must be considered to fully understand protein-ligand binding [19]. Under *in vivo* conditions, where drug and target are not in equilibrium, it is crucial to incorporate the kinetic rates of association and dissociation along with thermodynamic descriptions for a comprehensive understanding of the binding process [20].

1.2 Kinetics & Binding Kinetics

In chemistry, kinetics examines the speeds at which chemical reactions occur. It quantitatively describes chemical reaction rates, investigates the factors that influence them, and it aims to elucidate essential characteristics of reaction pathways [27]. To fully comprehend the specific binding established between biomolecules, it is essential to analyze structural, thermodynamic, and kinetic aspects in conjunction. Kinetic properties play a pivotal role in predicting drug efficacy and safety, enhancing the focus beyond the commonly prioritized binding affinity [20].

As anticipated, the protein-ligand interaction is based on the random coupling between the ligand and its biological receptor, where molecular diffusion plays a crucial role [28]. From a kinetic perspective, the protein-ligand binding process is analyzed as a two-step mechanism, where the first step is the formation of the so-called encounter complex, while the second step is the formation of chemical bonds [28]. Particularly, the ligand approaches its target via a random walk through the solvent during the diffusive phase, before the final formation of the encounter complex [4]. Molecular diffusion is an entropy-driven process, and the connection between entropy and diffusion is based on the foundational concepts of thermodynamics and statistical mechanics. Moreover, it depends on the original kinetic energy of the system.

As previously introduced, the system composed by solvent molecules, ligand and receptor is modeled by the simple thermodynamic system made up of solvent and solute molecules. In this context, the diffusion of solute components arises from the kinetic energy of the solute molecules, along with collisions between the protein, ligand, and solvent molecules. Solvent molecules move at varying speeds in random directions. Their collision with solute molecules may facilitate the association of the protein and ligand species and the final formation of non-covalent interactions.

Following the seminal works of vant'Hoff in 1884 [29] and Arrhenius in 1889 [30], theoretical models for calculating kinetic rate constants primarily rely on the energy barriers that separate the initial and final states of a system. Specifically, in the case of protein-ligand binding, the bound and unbound states. The highest energy barrier separating the initial and final states is associated with a system's state known as the transition state or the activated complex. The kinetic rate constants for the association k_{on} and dissociation k_{off} processes of the binary complex can be expressed as:

$$k_{on} \propto e^{\frac{-\Delta G_{on}^{\ddagger}}{RT}} \quad (1.16)$$

$$k_{off} \propto e^{\frac{-\Delta G_{off}^{\ddagger}}{RT}} \quad (1.17)$$

where the $\Delta G_{off}^{\ddagger}$ and $\Delta G_{on}^{\ddagger}$ are the free energy differences between the binary complex and the transition state, and the free energy differences between the dissociated states and the transition state, respectively. To clarify, the concept of the activated complex is intrinsic to the kinetic theory and is often used interchangeably with the transition state. However, it's important to distinguish between the two: the transition state refers to the specific configuration(s) with the highest energy, as defined by transition state theory, while the activated complex represents a collection of conformations near the transition state, as initially suggested by Arrhenius [27].

The relationship between reaction rates k , T , and activation energy is described by the Arrhenius rate law [30]:

$$k = Ae^{-\frac{E_a}{RT}} \quad (1.18)$$

or:

$$k = Ae^{-\frac{E_a}{k_B T}} \quad (1.19)$$

where R is the universal gas constant, A is the Arrhenius proportionality constant, and E_a represents the activation energy. The activation energy quantifies the energy barrier that must be overcome to escape stability energy basins for a reaction to proceed. The only difference between equation 1.18 and equation 1.19 is the unit of E_a , namely energy per mole (molar activation energy) and energy per molecule (unit as $k_B T$), respectively. Arrhenius formula shows a linear relationship between the logarithm of reaction rate and the inverse of the temperature, illustrating how the rate constant of a reaction depends on both temperature and activation energy. The pre-exponential Arrhenius factor is also called frequency factor, and it represents the frequency of collisions between reactant particles at standard concentration. It can be quantified both experimentally or numerically, and it depends on T because it is related to molecular collisions, which are a function of T as well. Considering two particles α and γ , the definition of A is different depending on the theoretical framework considered.

1. within the collision theory [31]:

$$A = d_{\alpha\gamma}^2 \sqrt{\frac{8k_B T}{\mu}} \quad (1.20)$$

where $d_{\alpha\gamma}$ is the collision radius between the two particles, and μ is the reduced mass of the system.

2. within the transition state theory [32]

$$A = \frac{k_B T}{h} \quad (1.21)$$

where h is the Planck's constant.

For both collision and transition state theories, A is responsive to temperature changes. In the collision theory framework, A is proportional to the square root of T (Eq. 1.20), whereas in the transition state scenario A is proportional to T (Eq.

1.21). However, in both cases, as the T increases, molecular motion accelerates, leading to higher kinetic energy. This increase in molecular speed results in more frequent collisions between molecules, thereby raising the collision frequency and the likelihood of overcoming the activation energy barrier required for reactions to occur. The units of the pre-exponential factor, as the unit of the rate constant, depends on the order of the reaction considered.

Collision and transition state theories are among the most important theories for chemical kinetics. However, the collision theory deals properly only with gases because it does not account for structural complexities of molecules, particularly biomolecules. In the collision theory scenario, chemical reactions requires collisions between reactants to occur, where molecules must be properly oriented and collision must have enough energy in order to result in product formation [31]. On the other hand, transition state theory provides a more accurate alternative with respect to the collision theory (and the simple Arrhenius equation). This new theory was devised by Eyring, Evans, and Polany, in order to include molecular mechanistic considerations [33]. Indeed, the transition state theory explains the rates of biochemical reactions in fluid environments, providing insight into the details of a reaction at the molecular scale. In 1935 Henry Eyring formulated the famous Eyring equation [34]. As Arrhenius, also the Eyring-Polanyi equation describes the kinetics rate constant k of a reaction as a function of the absolute T , but it involves a statistical frequency factor κ as following reported [34]:

$$k = \frac{\kappa k_B T}{h} \exp\left(-\frac{\Delta G^\ddagger}{RT}\right) \quad (1.22)$$

here, $\Delta G^\ddagger$ is the Gibbs free energy of activation, $h = 6.626\,070\,15 \times 10^{-34} \text{ J} \cdot \text{Hz}^{-1}$ is the Planck constant, and κ is the transmission coefficient. This latter term was introduced to account for the fact that the transition state does not always lead to product formation. It considers reactive trajectories that recross the transition state back and return to reactants without reaching the product state. As a result, κ value typically ranges between 0 and 1, but it is commonly assumed to be equal to 1. With κ equal to 1, the fraction of the flux passing through the transition state proceeds to the product without recrossing the transition state. To effectively

utilize transition state theory, a detailed description of the system’s potential energy surface is essential. Indeed, reaction rates can be analyzed by investigating activated complexes near the saddle point of this surface. However, accurately characterizing the structure of the transition state remains a challenging task.

1.2.1 Kinetics in Drug Discovery

In recent years, the importance of kinetics in drug discovery has increased, supported by evidence indicating that measuring unbinding kinetics can provide valuable insights into a drug’s *in vivo* efficacy. This insight shifted the focus beyond simple binding affinity to also consider the dynamic interactions between drugs and their targets [35, 36]. Indeed, drug efficacy not only depends on binding affinity but also on the time during which the drug remains associated with its biological target, known as residence time (RT). The underlying rationale is that a drug is efficacious only as long as it is bound to its target. Notably, the drug RT has emerged as an important aspect *in vivo* together with binding free energy [37, 38, 13, 39]. RT is defined as the inverse of the dissociation rate constant, following the simple equation:

$$RT = \frac{1}{k_{off}} \quad (1.23)$$

Additionally, under the reasonable assumption that k_{on} remains stable within a series of chemically related compounds sharing a similar scaffold, optimizing k_{off} can directly enhance binding affinity. This is due to the relationship reported in Eq. 1.6 (i.e., $K_d = k_{off}/k_{on}$). As a result, k_{off} is often considered a reasonable proxy for K_d , which is a traditionally sought quantity in drug discovery.

Many drugs currently on the market exhibit extended residence times, ranging from seconds to days, as molecules with prolonged RT often show improved *in vivo* efficacy. Moreover, RT not only affects the long-lasting pharmacodynamics of a molecule but also its pharmacokinetics [38]. Specifically, RT influences the drug’s absorption, metabolism, distribution, elimination, and toxicity (ADMET) profile, modulating key factors such as the duration of action and the compound toxicity [13, 19]. Consequently, predicting and optimizing RT is becoming increasingly crucial for the success of drug development, particularly during the early stages of

the drug discovery process [27, 13].

Various experimental techniques are available for studying binding kinetics. While these methods provide valuable information, they generally fall short of revealing atomic-level mechanistic details or delivering a comprehensive interaction map between ligands and their targets. Moreover, they offer limited structural insights into the transition states involved in ligand unbinding, which can be critical for designing molecules with improved reaction rates [27, 40, 41]. Theoretically, MD simulations are a highly suitable computational approach for studying and predicting binding kinetics [4]. However, standard MD simulations often struggle to sample all relevant conformational states within a practical computational time [42, 36, 43]. While ligand binding typically occurs within milliseconds, the unbinding process for lead candidates can take several minutes or longer, making unbiased simulations of unbinding events computationally expensive and impractical for routine use [4]. To address these limitations, a variety of enhanced sampling techniques have been developed. These advanced computational methods employ different strategies to achieve the extensive sampling needed to study these processes efficiently [35, 36, 41]. A more detailed discussion of these techniques is provided in section 2.4.

In conclusion, kinetic properties may serve as critical differentiators and predictors of drug efficacy and safety, in addition to the traditionally focused binding affinity. However the quantitative structure-kinetics relationship for modeling and ligand design is still not fully understood [20, 44]

Chapter 2

Theoretical background

Since the early 1970s, when biomolecular simulations first emerged, biological systems can now be effectively modeled through computer simulations. This is done by using microscopic models, where molecules and atoms are represented as interacting particles. The two primary models used to describe intermolecular and intramolecular interactions are quantum mechanics (QM) and **Molecular Mechanics** (MM). These models enable the calculation of the system's energy for any arrangement of atoms, and they track how energy evolves with particle movements [45]. Although the true nature of molecular interactions is quantum mechanical, many biologically relevant events occur under conditions where MM provides an accurate and computationally feasible description. QM, while more precise, is inherently more complex and computationally expensive, making it impractical for routine use in large-scale simulations.

Moreover, since molecules are composed of ensembles of atoms and often exist in multiple conformational states, **Statistical Mechanics** becomes the most appropriate approach to describe their behavior. Consequently, in biomolecular simulations, statistical mechanics is primarily applied within the framework of molecular mechanics, providing a balance between accuracy and computational efficiency.

Once the MM framework is chosen, the system's distribution can be sampled using various approaches, with **Molecular Dynamics** and Monte Carlo methods being

the most commonly used techniques.

2.1 Statistical mechanics

In classical mechanics, the analysis of a system involves a dual-level approach: the microscopic level and the macroscopic level. For microscopic systems, fluctuations are unavoidable during thermodynamics measurements because of the lack of control that we have in microscopic dynamics. On the other hand, the magnitude of fluctuations becomes very small for macroscopic systems. However, the macroscopic observables of a system are linked to its instantaneous microscopic states [46]. Statistical Mechanics seeks to understand the macroscopic properties of a system, which is made up of many particles, by starting from its microscopic description [47]. It combines two fundamental concepts: mechanics, which is deterministic, and probability theory, which introduces indeterminacy, providing a statistical approach for analyzing dynamical systems [48].

To fully specify the state of a system (composed by N particles) in the Cartesian space at each instant of time, a set of coordinates and conjugate momenta must be defined: $r_i(t) = [x_i(t), y_i(t), z_i(t)]$ and $p_i(t) = [p_i^x(t), p_i^y(t), p_i^z(t)]$ with $i = 1, \dots, N$. In this way, the system is described by the dynamical variables $\{r^N\}$ and $\{p^N\}$, where $\{r^N\}$ denotes an ensemble of N position vectors corresponding to $3*N$ scalar components. Similarly, $\{p^N\}$ represents the $3*N$ momenta vectors. The specific set of coordinates and momenta defines a point in the phase space, $\Gamma(r^N, p^N)$, namely a system's micro-state. The multidimensional phase space of the system Γ is the collection of all possible states that the system may have, where $\Gamma(r^N, p^N)$ is a particular point defined by $6*N$ variables. These $6*N$ variables completely describe the system's micro-state in the $6N$ -dimension hyperspace.

To determine the future behavior of the system, its evolution in time is described by $\{\Gamma(t)\} = \{r^N(t), p^N(t)\}$. This trajectory connects all the states in Γ visited by the system over time. When macroscopic conditions (e.g., fixed energy and volume) are imposed, the system explores only a manifold of phase space consisting of all the points that correspond to those specific values of the macroscopic variables.

Based on that, any interesting observable of the analysed system is expressed as function of $\Gamma(r^N, p^N)$ [49].

For analyzing dynamical systems, one of the most valuable observable is the system's total energy. According to the Hamiltonian formalism, the total energy of the system is described by the Hamiltonian function. The Hamiltonian for an N-particle system is:

$$\hat{H} = \hat{K} + \hat{U} = \sum_{i=1}^N \frac{p_i^2}{2m_i} + U(\{R\}) \quad (2.1)$$

where $\{R\}$ is the set of all particles' coordinates, and m_i is the mass of the i^{th} particle considered. As can be seen in Eq. 2.1, the Hamiltonian of a system is expressed as a sum of two different functions: the kinetic energy function ($\hat{K}$), which depends solely on the conjugate momenta, and the potential energy function ($\hat{U}$), which depends solely on the system's coordinates. In the Hamiltonian framework, the system's equations of motion are governed by Hamilton's equations:

$$\dot{r}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial r_i} \quad (2.2)$$

Solving Hamilton's equations provides the system's trajectory, which details the positions and momenta as functions of time. This trajectory describes the dynamic evolution of the system in phase space, determined by its initial conditions and the forces acting on it.

Complementary to the Hamiltonian is the Lagrangian formalism, which provides an alternative but equivalent description of the system's dynamics. The Lagrangian function, defined as the difference between kinetic and potential energy, is particularly useful in deriving the equations of motion in generalized coordinates via the principle of least action. While the Hamiltonian emphasizes energy conservation and phase space evolution, the Lagrangian offers a more direct route from physical principles to the system's trajectories. The Lagrangian formalism is especially valuable in more advanced contexts, such as constrained dynamics and the formulation of field theories, where its flexibility in handling generalized coordinates

becomes essential.

Statistical mechanics is founded on two fundamental assumptions: (i) given sufficient time, the system's initial conditions become irrelevant, and (ii) in a system at equilibrium, all micro-states have equal probability to be visited [47]. Using now the notation $\{r_i(t), p_i(t)\}$ to represent a specific micro-state of the system, one can get the average of a macroscopic observable A of interest. This is defined as the average of a microscopic operator $\hat{A}(r_i, p_i)$ over all accessible states. Consequently, a macroscopic property is defined (in terms of phase-space configurations) as the ensemble average:

$$\langle A \rangle = \int_{r_i, p_i} \hat{A}(r_i, p_i) \rho(r_i, p_i) dr_i^N dp_i^N \quad (2.3)$$

where $\rho(r_i, p_i)$ is the probability of the system being in state i . $\hat{A}(r_i, p_i)$ is the microscopic property corresponding to the specific micro-state, which is characterized by $3 * N$ r coordinates and $3 * N$ p momenta. While macroscopic properties remain time-independent at equilibrium, different points in the microscopic phase space are visited with diverse probabilities. The probability density associated with each micro-state is also time-independent, and, by definition, the total probability distribution over the entire phase space is normalized to one.

On the other hand, real experimental measurements yield time averages. Thus, for the same observable A , time averages involve averaging the microscopic function $\hat{A}(r_i, p_i)$ over the duration of the measurement:

$$\bar{A} = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \hat{A}\{r_i(t), p_i(t)\} dt \quad (2.4)$$

The relationship between these two perspectives is established by the ergodic hypothesis, which states:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \hat{A}\{r_i(t), p_i(t)\} dt = \int_{r_i, p_i} \hat{A}(r_i, p_i) \rho(r_i, p_i) dr_i^N dp_i^N \quad (2.5)$$

The ergodic hypothesis assumes that, for an isolated system and with infinite sampling time $t \rightarrow \infty$, the system will explore all possible micro-states. If the ergodic hypothesis holds, then time averages and ensemble averages will be equivalent (Eq.

2.5). In the ensemble average (Eq. 2.3), each micro-state contributes to the integral accordingly to its probability density. In contrast, the time average (Eq. 2.4) is taken in the infinite-time limit, assuming that the system explores all relevant micro-states during this period of time. However, the time required for a system to actually sample the entire available phase space is extremely large and increases nearly exponentially with the number of atoms in the system.

As emerged from this explanation, unlike classical mechanics which deals with the behavior of a single state, statistical mechanics uses the concept of a statistical ensemble. A statistical ensemble is a large collection of virtual, independent replicas of the same system, each one in a different possible micro-state. This ensemble provides a probability distribution over all possible states, enabling the prediction of macroscopic properties by considering the collective behavior of the entire system [47].

2.1.1 Statistical ensembles

In the statistical mechanics context, the behavior of a system is examined by considering all the possible micro-states it can occupy, which share specific macroscopic properties as volume (V), energy (E), or number of particles (N). The possible micro-states that the system can occupy are collectively referred as the system's ensemble of states [22, 46]. Different types of ensembles are used to describe different thermodynamic conditions. However, it is important to note that all ensembles are equivalent within the thermodynamic limit [47].

Microcanonical ensemble

The microcanonical ensemble describes a system under conditions of constant E , N , and V ; i.e. a system with a constant and known energy E , a constant number of particles N and a fixed known volume V . The microcanonical ensemble includes all possible states that are compatible with such NVE setting. The density of state (Ω), or the microcanonical partition function, counts the total number of

micro-states of the system consistent with each macro-state:

$$\Omega(N, V, E) = \frac{1}{h^{3N} N!} \int \int_V \delta[H(r, p) - E] dr dp \quad (2.6)$$

where δ is the Dirac function, r and p represent the $\{r^N\}$ and $\{p^N\}$ coordinates and conjugate momenta of all particles, and H is the Hamiltonian of the system. The integrals are done over the entire $6N$ -dimensional space, the term $N!$ accounts for the indistinguishable states, and the Dirac delta function selects only the states having specific energy E . The $\frac{1}{h^{3N}}$ is a simple proportionality factor where h^{3N} is the quantized volume of a micro-state (the Planck constant measures the elementary volume in the phase space). The density of states interpretation leads to the fact that the quantity $\Omega(E)dE$ measures the number of states having energy $E \pm dE/2$. Moreover, the microscopic density of states is directly related to the macroscopic entropy of the system through Boltzmann's equation (i.e., $S = k_B \ln \Omega$). In the microcanonical ensemble, quantities such as temperature are derived properties. These properties are computed by counting the number of micro-states available at fixed energy.

Canonical ensemble

The canonical or isochoric-isothermal ensemble describes a statistical system under conditions of constant N , V , and T . The probability density of this ensemble is defined as:

$$\rho_{NVT}(r_i, p_i) = \frac{1}{h^{3N} N!} \frac{e^{-\beta H(r_i, p_i)}}{Q_{NVT}} \quad (2.7)$$

where the inverse thermal energy β is equal to $\frac{1}{k_B T}$. An ensemble of micro-states obeying to this probability density is said to be Boltzmann-distributed; i.e., lower-energy states are more probable than higher-energy ones, with the distribution shaped by the system's temperature. By definition, the integral of the probability distribution function over the entire phase space must sum to one. In equation 2.7, the canonical partition function Q_{NVT} represents the required normalized factor:

$$Q_{NVT} = \frac{1}{h^{3N} N!} \int_{r_i, p_i} e^{-\beta H(r_i, p_i)} dr_i^N dp_i^N \quad (2.8)$$

In the partition function equation (Eq. 2.8), the values of the energy of micro-states are summed over and they can vary, meaning that energy fluctuations are non-zero. Q_{NVT} depends on the temperature through the inverse thermal energy β . Since the total energy of the system can be decomposed into the kinetic energy term (which depends on momenta and is analytically computable) and the potential energy term (which depends on coordinates), the partition function can be separated into two distinct components: one corresponding to the kinetic or ideal gas part, and the other to the potential or excess part. The excess part is the configurational integral:

$$Z_{NVT} = \int dr^N e^{-\beta U_{NVT}(r^N)} \quad (2.9)$$

The partition function plays a crucial role in statistical mechanics. Once the partition function is exactly determined, it enables the straightforward derivation of any thermodynamic property of the system. However, in almost all cases, Q_{NVT} can not be computed from its equation practically. This is due to the configurational integral, whereas the kinetic component is analytically integrable.

In contrast with the microcanonical ensemble, here T is a fixed quantity, while the average energy becomes a derived quantity. In the context of this ensemble, the Helmholtz free energy is a thermodynamic function determined as:

$$F(N, V, T) = -k_B T \ln Q_{NVT} \quad (2.10)$$

$$F(N, V, T) = U(N, V, T) - TS(N, V, T) \quad (2.11)$$

where $U(N, V, T)$ is the thermodynamic internal energy, while $S(N, V, T)$ is the entropy of the system. As for the other statistical ensembles, by using standard thermodynamic definitions and relations, all thermodynamic properties can be derived.

Isothermal isobaric ensemble

The isothermal isobaric ensemble describes a statistical system under conditions of constant N , P , and T . The isothermal-isobaric partition function is:

$$Q_{NPT} = \frac{1}{h^{3N} N!} \int_0^\infty dV \int_{r_i, p_i} e^{-\beta(H(r_i, p_i) + PV)} dr_i^N dp_i^N \quad (2.12)$$

where the integral over micro-states includes micro-states with different volumes. The factor $-\beta PV$ weights each volume according to the probability of observing that volume under a given P .

In the context of this ensemble, the Gibbs free energy is defined as:

$$G(N, P, T) = -k_B T \ln Q_{NPT} \quad (2.13)$$

This ensemble is widely used to maintain realistic biological conditions, allowing the system to adjust its volume dynamically to match the applied pressure. By using standard thermodynamic relations, other thermodynamic quantities can be derived. We should also mention that given the very small value of PV in atomistic simulations the Helmholtz and Gibbs free energy gives almost identical estimations.

Grand canonical ensemble

The last statistical ensemble is the grand canonical ensemble, which describes a statistical system under conditions of constant T , μ , and V ; where μ is the specified chemical potential of the system. Hence, the number of particles becomes a variable. The grand canonical partition function is:

$$Q_{\mu VT} = \sum_{N=0}^{\infty} \frac{\exp[N\beta\mu]}{h^{3N} N!} \int e^{-\beta H(r_i, p_i)} dr_i^N dp_i^N \quad (2.14)$$

The grand-potential is:

$$\Omega(\mu VT) = -k_B T \log Q_{\mu VT} \quad (2.15)$$

As for the other statistical ensembles, by using standard thermodynamic relations, the other thermodynamic quantities can be derived.

2.1.2 Molecular systems

So far, the statistical mechanics equations have been introduced for a simplified system consisting of a single component made of isotropic particles. However, when applied to molecular systems, the complexity increases as we are often dealing with multi-component systems that are mixtures of molecular fluids or solids. In this context, molecules of the same species are considered indistinguishable, while the atoms within each molecule are treated as distinguishable. This distinction is due to the fact that covalent bonds, which are fixed in the MM model, maintain the relative positions of the atoms.

Therefore, given the complexity of real systems with numerous interacting particles, it is impractical to sample the entire configuration space comprehensively. Only configurations with the highest probabilities, which contribute significantly to the ensemble average $\langle A \rangle$ (Eq. 2.3), are typically sampled. The most representative configurations are those with the lowest energy and the highest probability. To generate molecular ensembles following the statistical Boltzmann distribution, computer simulations are commonly used. In particular, in the realm of CADD, MD simulations and related techniques are on the verge of becoming fundamental computational tools [11]. However, it is important to note that when regions of high probability in configuration space are separated by significant energy barriers, standard MD techniques may struggle to adequately sample all relevant configurations. Solutions to this issue, referred to as quasi-non-ergodicity, require specialized approaches [11]. Enhanced sampling methods, designed to overcome these limitations, will be discussed in detail later in this chapter (Section 2.4).

Computer simulations provide detailed information of the system at the microscopic level, capturing atomic or molecular positions and velocities. MD simulations produce microscopic structural ensembles, with statistical weights determined by the chosen theoretical model. It is through statistical mechanics that these microscopic data are translated into meaningful macroscopic properties, allowing direct comparison with experimental measurements [49].

2.2 Molecular Mechanics

Molecular mechanics is an empirical computational method that applies the principles of classical mechanics to simulate molecular systems, operating under the assumption that the Born-Oppenheimer approximation holds [45]. Unlike quantum theory, which offers a more accurate representation of real molecules, the molecular mechanics approach does not explicitly account for electrons. Instead, it implicitly incorporates the effects of different electronic configurations of atoms through the use of force field parameters and tracks the nuclei motions only [50]. Specifically, atoms are treated as elementary units within molecular systems, having the movements of electrons and nuclei decoupled. This separation is a direct consequence of the Born-Oppenheimer adiabatic approximation [51].

In 1927, Born and Oppenheimer proved that the Schrödinger equation for a molecular system can be separated into two distinct equations, namely the electronic equation and the nuclear term. Starting from the Hamiltonian (now operator) of the system, with both electron and nuclei degrees of freedom combined and no relativistic effects included (such as spin-orbit coupling), the system's total energy is:

$$\hat{H} = \hat{K}_{nucl} + \hat{K}_{el} + \hat{U}_{nucl-nucl} + \hat{U}_{el-el} + \hat{U}_{nucl-el} \quad (2.16)$$

where the first 2 terms correspond to the kinetic energy of the nuclei and electrons, respectively. The last 3 terms represent, in order, the potential energy for the Coulomb interactions: the repulsion between two nuclei, the mutual repulsion between electrons, and the attraction between nuclei and electrons. The evolution of the system over time is determined by solving the Schrödinger equation, to be either time-dependent or time-independent:

- Time-independent Schrödinger equation:

$$\hat{H}(\hat{r}, \hat{R})\Psi(\hat{r}, \hat{R}) = E\Psi(\hat{r}, \hat{R}) \quad (2.17)$$

- Time-dependent Schrödinger equation:

$$i\hbar \frac{\partial \Psi(\hat{r}, \hat{R}|t)}{\partial t} = \hat{H}(\hat{r}, \hat{R}) \Psi(\hat{r}, \hat{R}|t) \quad (2.18)$$

Ideally, the time-dependent Schrödinger equation can predict all properties of any molecule [22].

In these equations, lowercase variables refer to electrons while uppercase ones refer to nuclei. $\Psi(\hat{r}, \hat{R})$ is the total vibronic wave function, and refers to the state that combines the electronic and the vibrational (nuclear) components of a molecule. Solving exactly the molecular Hamiltonian for this combined system is computationally infeasible for most of the real-world problems, mostly because of the complexity of the interactions between nuclei and electrons. Therefore, the Born-Oppenheimer approximation is used to make this problem more tractable, treating electronic and nuclear motions separately, and reducing the computational difficulty. The Hamiltonian of the system can be now divided into separate electronic and nuclear terms:

$$\hat{H} = \hat{H}_{nucl} + \hat{H}_{el} \quad (2.19)$$

where $\hat{H}_{el}$ is the electronic Hamiltonian:

$$\hat{H}_{el} = \hat{K}_{el} + \hat{U}_{el-el} + \hat{U}_{nucl-el} \quad (2.20)$$

This decoupling arises because nuclei are significantly more massive than electrons ($M \gg m$). As a result, the kinetic energy of the nuclei is relatively small compared to the other terms in the Hamiltonian. This justifies the assumption that nuclei evolve more slowly than electrons. From a pictorial perspective, the nuclei experience the electrons as a rapidly adapting cloud of negative charge that binds them in fixed relative positions. Meanwhile, the electrons move and adjust almost instantaneously and adiabatically, in response to the slower motion of the nuclei. Thus, the Hamiltonian of the molecular system can be expressed as a function of the nuclear coordinates, with the fast motion of the electrons effectively averaged out. This approximation allows to focus on the slower nuclear dynamics while treating the electronic structure as adapting instantaneously to the nuclear positions.

For every fixed set of nuclear coordinates, the focus shifts to solving the electronic problem yielding, in principle, an infinite set of eigenvalues and corresponding eigenvectors. The wave function for this electronic state $\Psi_{el}(\hat{r}, R)$ depends only on $\hat{r}$, and is defined as:

$$\hat{H}_{el}\Psi_{el}(\hat{r}, R) = E_{el}(R)\Psi_{el}(\hat{r}, R) \quad (2.21)$$

By solving the equations for electronic wavefunctions and iterating over different values of R , which now is treated as a parameter, we can determine the electronic potentials on which the nuclei move. Furthermore, the assumption that electrons evolve rapidly implies that electrons can relax to their ground state much faster than the timescale of nuclear motion. As a result, the only significant electronic state is the ground state. In simple words, the Born-Oppenheimer approximation enables the description of the system’s potential energy solely as a function of the atomic nuclei coordinates.

In MM approaches, the system’s total energy is calculated by combining kinetic and potential energy. The kinetic energy term can be solved analytically, while the potential energy is determined based on the atomic nuclei’s positions, by using predefined mathematical models known as force fields (FF). This potential energy function is additive, being the sum of bonded and non-bonded energy terms. The MM approach is versatile, enabling the modelling of molecular systems across a wide range of sizes and complexities: from small molecules to large biological structures containing thousands or even millions of atoms. The MM method itself is frequently referred to as the force field method [52]. A more detailed explanation of the empirical force fields used in MM simulations can be found in Section 2.3.2.

2.3 Molecular Dynamics

Molecular Dynamics is a computational technique that simulates atom motions along time on computers at finite temperatures. As mentioned earlier, macroscopic properties observed in experiments are not direct measurements; rather, they are averages derived from the behavior of Avogadro numbers of molecules that form a statistical ensemble. The two prevalent methods for generating equilibrium en-

sembles are Monte Carlo and Molecular Dynamics simulations. The latter offers the added benefit of accurately capturing the kinetics of nonequilibrium processes with temporal resolution.

Recent advancements in the physico-chemical analysis of biological processes underscore the growing significance of MD simulations [53], which in turn rely on empirical FFs. These simulations are increasingly important in enhancing the understanding of biomolecular structures and their functions. By providing valuable insights into the dynamics of biological systems, MD simulations aid in interpreting experimental data and expedite discoveries in molecular biology [53, 11].

The ability of MD to calculate equilibrium properties is based on the principles of the ergodic hypothesis we previously discussed (Eq. 2.5) [22]. The aim of MD simulations is to investigate the relationships between the macroscopic properties of a molecular system and its atomic features. More specifically, the objective is to quantitatively evaluate macroscopic quantities that are derived from microscopic models [53]. The essential idea behind a MD simulation is to predict how atoms in a molecular system evolves over time by integrating Newton’s laws of motion for each atom. Given a molecular system, starting from the initial positions of its atoms (i.e., the starting conformation), the forces acting on each particle are calculated using principles of classical mechanics (i.e., Newtonian mechanics). At each simulation step, the forces on all atoms are repeatedly computed. Then, forces are used to update the atoms’ positions and velocities, tracking atoms’ motion over time [46].

In classical physics, the motion of a particle with mass m_i along a given coordinate x_i is governed by Newton’s second law:

$$\frac{\partial^2 x_i}{\partial t^2} = \frac{F_{x_i}}{m_i} \quad (2.22)$$

which can be expanded to cover all the spatial dimensions in a realistic molecular system. In this equation (Eq. 2.22), F_{x_i} represents the force acting on the individual i -th particle along a specific direction. This force is the resultant of the forces exerted on the atom by all the other atoms; the more the neighbor atoms are dis-

tant, the more negligible their contribution becomes. This observation is crucial, as it allows for significant computational time savings by approximating or neglecting the influence of distant atoms, thereby optimizing the calculation process [54].

In other words, for a molecular system composed of N interacting atoms, MD simulations solve the classical Newton’s equations by calculating the forces acting on each atom as the negative derivatives of the system’s potential energy function [54]. Therefore, for a system with N particles and represented as a set of atomic coordinates and conjugate momenta, equations are:

$$m_i \frac{\partial^2 x_i}{\partial t^2} = F_{x_i} \quad (2.23)$$

$$F_{x_i} = \frac{dp_i}{dt} = -\frac{\partial U(x_i)}{\partial x_i} \quad (2.24)$$

with $i = 1, \dots, N$. Actually, the force is defined as the negative derivative of the energy with respect to the coordinates. Since the kinetic energy is independent of positions, the force is equivalent to the negative derivative of the potential energy with respect to the coordinates, as in Eq. 2.24. Integrating these equations numerically requires knowing analytically the molecule’s potential function (or potential energy surface, PES).

The most principled way to obtain the molecular PES is by computing the energy through the solution of Schrödinger’s equation, and then determining the forces from derivatives. However, using the Born-Oppenheimer approximation, this process is simplified by assuming that the motion of the nuclei is influenced only by the mean-field of electrons. Thus, the focus is on the positions and forces acting on the nuclei, while the electronic effects are captured through the force field parameters under specific approximations (i.e., electrons adiabatically follow nuclear motion, and nuclei evolve on a single Born-Oppenheimer PES). As a result, the forces are computed using empirical FFs, which are designed to approximate the system’s PES enabling the evaluation of atomic interactions. By definition, the FF is a theoretical model composed of mathematical functions, and the corresponding constants, that describe all interatomic and intermolecular interactions.

To compute the trajectory of a system with N particles, a system of $3N$ second-order differential equations must be solved. To overcome this problem, these equations are solved iteratively using integration techniques (i.e., equations are solved simultaneously in small time steps), which are based on the time discretization of the evolution of the system. Starting from a suitable initial conformation, and given the position and its derivatives at time t , dynamic quantities are estimated at $t + dt$, where dt is a small finite time increment. Integration methods are iterative methods, and they are typically derived from Taylor expansions of atomic positions and other dynamic variables with respect to time. Many algorithms and integration rules, known as integrators, are available for solving the equations of motion. Two of the most widely used are (i) the Verlet Algorithm [55], and (ii) the Leapfrog Algorithm [56].

2.3.1 Integration Algorithm

The integration algorithm for a MD simulation starts by assigning random velocities to the system's atoms, which are selected from a Maxwell-Boltzmann distribution at a given T . This is necessary only when the system has an unknown initial velocities state. Then, the algorithm computes the forces exerted on every atom to successively calculate the atoms' acceleration. This information is combined with the positions and velocities of the atoms at time t to predict a new set of velocities and positions at time $t + dt$. At $t + dt$ a new collection of forces is computed, and the procedure keeps going repeatedly until the desired simulation time is reached. Thus, for the numerical integration of the equations of motion, finite difference methods are employed, where the choice of an appropriate time step size is crucial for accurate results [22].

To guarantee numerical stability, the overall energy conservation, and the continuity in the atomic forces, the time step dt must be very small. From this prospective, the smaller the dt , the better the approximation and the results accuracy. However, small dt means that the number of steps needed to sample completely the molecular space increases, requiring high computational resources and time. Conversely, larger time steps offer computational benefits but can lead to issues during calcu-

lations, such as atomic clashes, atomic overlaps, or instances where particles fail to interact properly (particles overlooking each other). Therefore, the correct balance for the time step selection must be found. To address this issue effectively, the system temperature T is typically considered: as the T increases, atomic velocities rise, requiring a smaller time step [57]. Additionally, the time step is generally chosen to be orders of magnitude smaller than the significant molecular movements, which leads to femtosecond units (10^{-15} seconds) commonly employed in biomolecular simulations. More precisely, the Nyquist sampling theorem states that the simulations' time step should be no more than half the period of the fastest dynamics, with a recommendation to use an even smaller time step, typically between 0.1 and 0.2 times the shortest period [45]. Time steps can vary from 0.25 femtoseconds for lighter nuclei, like hydrogen, to 2 femtoseconds or more for heavier nuclei. For instance, the vibrational frequency of a C-H bond is around 3000 cm^{-1} , which corresponds to a period of approximately $8.99 \times 10^{13}\text{ Hz}$ (11 femtoseconds). Thus, for a biological systems, where the C-H stretching mode is the highest frequency oscillator with a period of about 10/11 femtoseconds, the time step is typically set to 1 femtoseconds.

To accelerate simulation computational times increasing the time step without compromising accuracy would be the ideal route. The combination of MD with constraints methods makes this possible without incurring in the previously mentioned issues. Practically, instead of integrating fast degrees of freedom, such as bond vibrations, these can be fixed at their equilibrium values, allowing for larger time steps as the corresponding fast vibrational frequency is now no more present in the system. While the constrained system is not statistically equivalent to the original flexible one, it has a minimal effect on average geometries due to small bond displacements [45].

In MD, constraints are incorporated into the equations of motion through the use of Lagrange multipliers. These multipliers are calculated based on the equations representing the constraints. The method which introduced such concept in MD is called SHAKE. It was developed for satisfying bond geometry constraints during numerical simulations, ensuring that these constraints are precisely fulfilled at every

time step [58]. The corresponding algorithm works by solving a set of Lagrange multipliers within the constrained equations of motion. The SHAKE approach has propelled many variants of it, being the conceptual approach identical. Variations include the LINear Constraint Solver (LINCS) and the SETTLE algorithms:

- the LINCS algorithm resets bond lengths to their correct values after an unconstrained MD step. The algorithm is inherently stable, and it is faster (3 to 4 times faster) and more efficient than the first iterative proposal of the SHAKE approach. It is particularly suitable for parallel computing environments [59].
- the SETTLE algorithm is specifically designed for resetting the positions and velocities of rigid water models to satisfy the holonomic constraints. It is very efficient for small molecules as water molecules [60].

Generally, the SHAKE algorithm and its variants are used to constrain high-frequency vibrations (typically bonds involving hydrogen atoms), allowing for a time step of 2 femtoseconds in standard MD simulations [22]. The choice of time step and the use of constraint algorithms are crucial factors in balancing the accuracy and efficiency of MD simulations. Finally, the simulation ends when enough steps have been integrated, reaching the desired simulation time

In this thesis, all MD simulations were performed using the LINCS algorithm with a 2-femtosecond time step, which was adequate for capturing the relevant dynamics of the biological molecules under investigation. However, when flexible water models were used (Chapter 5), a shorter, 1 fs, time step was necessary.

Verlet Algorithm

The Verlet algorithm is derived from the Taylor expansion of the atomic positions. It starts from the forward and backward Taylor series of the atomic positions at time t :

$$r(t + \delta t) = r(t) + v(t)\delta t + \frac{a(t)\delta t^2}{2} + \frac{b(t)\delta t^3}{3} + o\delta t^4 \quad (2.25)$$

$$r(t - \delta t) = r(t) - v(t)\delta t + \frac{a(t)\delta t^2}{2} + \frac{b(t)\delta t^3}{3} + o\delta t^4 \quad (2.26)$$

By summing these 2 expansions (truncated at the second order), the Verlet algorithm is obtained:

$$r(t + \delta t) = 2r(t) - r(t - \delta t) + a(t)\delta t^2 \quad (2.27)$$

Here, $r(t)$ is the position vector, v is the velocity $\dot{r}(t)$, a is the acceleration vector $\ddot{r}(t)$, dt is the integration time step, and b is the third derivative of the position with respect to the time $\dddot{r}(t)$. Despite the relative simplicity of the Verlet algorithm, it presents some limitations. Specifically, the velocity is not explicitly present in the formula, but it can be computed only *a posteriori*. Moreover, possible numerical errors may emerge during the calculations linked to the presence of a sum between different order terms in the calculations. Additionally, the algorithm requires a prior knowledge of the previous state, necessitating the use of an alternative integration methods to initialize its first step.

Leap-frog algorithm

The Leap-frog algorithm uses the atomic positions r at time t , and the velocities v at time $t - \frac{1}{2}\delta t$. The positions at time t are defined by the equations below, and they are used to compute the forces $F(t)$, which in turn are used to update positions and velocities:

$$r(t + \delta t) = r(t) + v(t + \frac{1}{2}\delta t)\delta t \quad (2.28)$$

$$v(t + \frac{1}{2}\delta t) = v(t - \frac{1}{2}\delta t) + a(t)\delta t \quad (2.29)$$

This algorithm represents an enhancement over the Verlet algorithm by eliminating the mixing of different order terms and enabling explicit treatment of velocity. However, due to the use of half-step velocities, there can be a misalignment between the velocity and position, resulting in a lag in their correspondence.

In all MD simulations performed in this thesis, Newton's equations of motion were integrated using the leap-frog algorithm.

2.3.2 Interactions treatment and Force Fields

The time-dependent Schrödinger equation theoretically allows for the prediction of all molecular properties for any molecule. However, as the number of particles in the system increases beyond a few, approximations become necessary to make calculations feasible. It is important to emphasize that even for simple systems, solutions of the Schrödinger equation, while highly accurate, remain approximate in nature. For complex systems, such as those investigated in this Thesis, these approximations accumulate, and the resulting solutions are often far from being fully reliable [22].

In MD simulations, the molecular system is represented as a collection of charged particles (atoms) connected by spring-like bonds, where interatomic and intermolecular forces are approximated by an analytical expression based on classical mechanics. MD simulations use empirical parameterizations of the molecular models, with parameter sets derived from the fitting of experimental data or the results from QM calculations. These parameters, known as FFs, are essential for calculating interactions and evaluating the free energy surface of the system. The FF comprises both the equations used to determine potential energy and forces (from particle coordinates), and the collection of the associated parameters. In most cases, the approximations made by FFs are effective. However, they fail to capture quantum effects, such as bond formation or breaking.

Force Fields functional form

All-atom FFs are interatomic potentials expressed as functions of atomic coordinates. They provide parameters specific to different atom types and depict the energy landscape of a given molecular system. In this context, an atom type refers to a categorization of atoms based on their chemical identity and the associated parameters that define their behavior within the simulation. Biologically-based FFs may have different flavors, but the basic functional form of potential energy is commonly divided into two contributes. Bonded interaction terms represent the interactions between directly connected (chemical bonds) atoms, while non-bonded terms represent fields induced, long and short terms interactions [22].

Molecules are represented as collections of atoms (each one with its type), where bonded and non-bonded interactions are treated differently depending on the specific FF. While the specific decomposition of the energy terms depends on the particular FF, the general form of the potential energy of the system for each point in configuration space $U(\{r^N\})$ is simply a sum of the two separate energy terms:

$$U(\{r^N\}) = U_b(\{r^N\}) + U_{nb}(\{r^N\}) \quad (2.30)$$

where $U_b(\{r^N\})$ and $U_{nb}(\{r^N\})$ represent the bonded and non-bonded energy terms, respectively. By using FFs based on two-body potentials for non-bonded interactions, only the interactions between pairs of atoms are evaluated. Therefore, the energy evaluation time, in a naive implementation, scales quadratically with the number of atoms for this energy term. A graphical representation of this empirical potential energy function is shown in Figure 2.1. Here, bonded interactions involve atoms that are directly connected or linked through a common atom, whereas non-bonded interactions occur between atoms that do not share a direct bond.

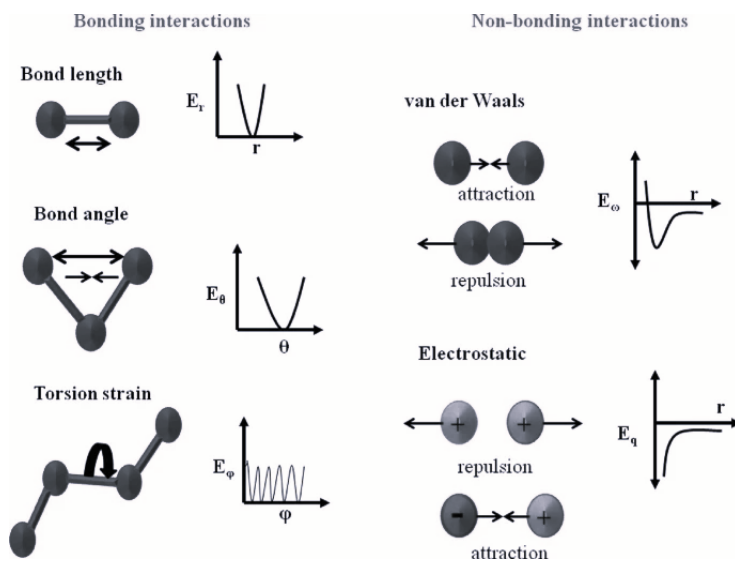


Figure 2.1: Schematic representation of the empirical potential energy function [61]. The energy expression and its parameters constitute the Force Field.

Bonding interactions are determined by the molecular topology, which is defined by the arrangement of covalent bonds between atoms within molecules. They

include bond stretching, angle bending and torsional potentials for bond rotations. Usually, bonded interactions represent up to four-body interactions. The standard representation of these terms is:

$$U_b = U_{stretching} + U_{bending} + U_{torsion} \quad (2.31)$$

Particularly, in the same order and explicitly:

$$\begin{aligned} U_b(\{r^N\}) = & \frac{1}{2} \sum_{ij} k_{ij} [r_{ij} - \bar{r}_{ij}]^2 + \frac{1}{2} \sum_{ijk} k_{ijk} [\theta_{ijk} - \bar{\theta}_{ijk}]^2 \\ & + \frac{1}{2} \sum_{ijkl} k_{ijkl} [1 + \cos(n\phi_{ijkl} - \bar{\phi}_{ijkl})]^2 \end{aligned} \quad (2.32)$$

In this context, i, j, k, l represent atoms connected by consecutive covalent bonds. The parameters $\bar{r}_{ij}, \bar{\theta}_{ijk}, \bar{\phi}_{ijkl}$ denote the reference values for bond lengths, bending and dihedral angles, respectively while, $k_{ij}, k_{ijk}, k_{ijkl}$ are the corresponding force constants. The non-negative integer n represents the periodicity of the torsion potential, which typically is $n = 2$ or $n = 3$. This term might be expressed with different functional forms depending on the atoms taken into account.

The dominant energy contribution in equations 2.31 and 2.32 comes from the stretching terms, which describe interactions between covalently bonded atoms. These interactions may be modeled effectively using the Morse potential. However, the Morse potential approach is rarely used because its exponential term is computationally demanding, and it requires to define three parameters. Under typical conditions, the energy required for bond compression or elongation is substantial, so atoms usually oscillate near their stable bond length. For small deviations (from the equilibrium bond distance), the stretching behavior can be approximated using a Taylor series expansion, keeping only its quadratic term. Consequently, $U_{stretching}$ is approximated by a harmonic potential, which models oscillations around the equilibrium bond length with a specific force constant. However, this model is reliable only under standard conditions and is not valid to study bond formation or breaking.

Similar to the stretching potential, the angle bending contribution describes the

energy associated with the bending of the angle between three bonded atoms. It represents the oscillations around a specific equilibrium angle, and it is modeled as a harmonic function with its specific force constant. This energy term is typically much higher than thermal energy.

Lastly, the torsion energy term describes the energy associated with the variation of dihedral angles. It applies to four sequentially bonded atoms, and it is defined by the angle of rotation around the covalent bond connecting the two middle atoms. In case of low energy barriers to internal rotation, torsional angles can undergo significant variations. As a result, the torsion energy $U_{torsion}$ is modeled using a periodic cosine expansion with multiplicity n . This model commonly relies on a short Fourier sum, with the number of included terms depending on the complexity of the potential and the required accuracy. Often, n is taken up to 4.

Additionally, extra terms such as out-of-plane "improper torsion" or coupling potentials may also be incorporated in the FF functional form. All the FF parameters (Eq. 2.32) are carefully chosen to replicate molecular properties that are measured experimentally or calculated using QM methods.

On the other hand, **non-bonding interactions** encompass both non-electrostatic and electrostatic interactions between pairs of atoms. They are mainly pair-additive, and they include Coulomb forces, short-range repulsion due to Pauli exclusion, and dispersion (van der Waals) forces. These interactions are represented as:

$$U_{nb} = U_{electrostatic} + U_{non-electrostatic} \quad (2.33)$$

Particularly, with the same order:

$$U_{nb}(\{r^N\}) = \frac{1}{4\pi\epsilon_0} \sum_{i \neq j} \frac{q_i q_j}{R_{ij}} + 4 \sum_{i \neq j} \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{R_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}} \right)^6 \right] \quad (2.34)$$

Here, the first term is the classical Coulomb potential, where q_i, q_j are the atomic charges, R_{ij} is the distance between them, and ϵ_0 is the vacuum's electric permittivity (in MD typically the interaction with the solvent is explicit, hence the vacuum constant presence). Atomic point charges are assigned to the positions of atomic

nuclei, and they are derived from QM methods to approximate the electrostatic potential surrounding the molecule. Commonly, Coulomb interactions are modeled using the rigid ion approximation, although there is a growing recognition of the importance of polarization effects. Additionally, pairs of atoms that are separated by one or two consecutive bonds are typically excluded, while the contributions from pairs separated by three consecutive bonds may be diminished (often by a factor of two).

The second term in Eq. 2.33 represents the non-electrostatic contribution, which includes short-range repulsion due to Pauli exclusion and dispersion forces. These interactions are typically modeled using the Lennard-Jones (LJ) potential, which is the last term in Eq. 2.34. Here, R_{ij} is the distance between the pair of atoms, while σ_{ij} and ϵ_{ij} are the length and energy scales of the potential. ϵ_{ij} denotes the well depth, which indicates the strength of the interaction. σ_{ij} defines the distance at which the intermolecular potential between the two particles becomes zero. The LJ interactions between different atom types are calculated by combining their respective LJ parameters (for instance σ_i and σ_j). Various FFs employ different combination rules, with arithmetic and geometric means being the two most common approaches. These combining rules help reduce the number of parameters needed for each possible combination of different atom types. Moreover, the exponent of 12 approximates the strong Pauli repulsion that occurs when electron orbitals overlap, whereas the exponent of 6 represents the weaker attractive forces arising from dynamically induced dipoles in the valence orbitals. While the attractive term is physically realistic (as London dispersion forces have a distance-dependent nature), the repulsive term is a simplified approximation of exponentially decaying repulsive interactions. This steepness of the repulsive part can lead to an overestimation of the pressure within the system.

In this scenario, the Lennard-Jones potential represents the short-range molecular interactions (i.e., the potential decreases faster than r^{-d} , where r is the distance between the particles and d is their dimension), while the Coulomb interaction is long-ranged and represents a key force for biological interactions.

In organic chemistry, covalent bonds exhibit a significant degree of transferabil-

ity between different macromolecules, facilitating the development of FFs valid for large groups of compounds. This extensive applicability simplifies the exploration of the diverse chemical space. Widely used FF parameterizations are Amber [62, 63], OPLS [64], CHARMM [65, 66], and Gromos [67]. Over time, these FFs have evolved into more refined versions, or they have been adapted into specialized subsets aimed at specific classes of molecules. In addition to macromolecules like proteins, also small molecules play a crucial role in computational biology, particularly in CADD. This has created a demand for more comprehensive and versatile FFs to accurately model small molecules that interact with proteins. These generic FFs include innovative parameters for a wide range of organic molecules that were previously not considered in theoretical models focused on macromolecules.

Amber Force Fields

In 1984, Weiner and colleagues introduced a united-atom FF in the Amber software, aimed at simulating proteins and nucleic acids [68]. This model used equilibrium bond lengths and angle parameters based on crystal structure data, while dihedral parameters were adjusted to align with torsional barriers obtained from both experimental findings and QM methods. The atomic charges were assigned using the Hartree-Fock STO-3G approach, and van der Waals parameters were taken from the work of Hagler *et al.* [69]. By 1986, an all-atom variant of the Amber FF was developed. With the rise in computational power, more refined parametrizations of both dihedral angles and partial charges were achieved, leading to enhanced versions of the Amber FFs [70, 71, 72, 63].

Additionally, to specifically parametrize and modeling small molecules, the General Amber Force Field (GAFF) was developed [73]. GAFF is a widely used, complete (no missing-parameters), and comprehensive FF. This FF is fully compatible with the protein Amber force field hence it shares completely the potential function form with it. Compared to other FFs, the atom types of GAFF are more general covering a wider range of the organic chemical space. The complete list of these atom types is available in literature [73]. The parameters of GAFF are well-suited for a broad range of organic compounds, and they undergo regular revisions to

improve accuracy. The latest generation, GAFF2 [74], includes updated parameters for enhanced modeling precision. Both GAFF and GAFF2 have demonstrated reliable performance in numerous scientific studies and are compatible with many standard FFs used for macromolecules. Given the heterogeneity of chemical compounds, parameters are assigned exactly when possible or by analogy. From one side this gives a remarkable flexibility, from the other the assigned parameters in some cases could be suboptimal. In GAFF small molecules parameterizations do not include atomic charges, requiring calculation via semi-empirical methods or ab initio techniques (they must be applied individually for each specific case). In the GAFF context, the HF/6-31G* Restrained Electrostatic Potential (RESP) charge method [75] offers the highest accuracy but, for large datasets, AM1-BCC (Austin Model 1 semi-empirical QM calculations with Bond Charge Corrections) charge method [76] provides a computationally very efficient alternative. AM1-BCC charges are specifically optimized to approximate HF/6-31G* RESP results, achieving reasonable accuracy within significantly reduced computational time. A detailed explanation of these charge methods is provided in the following section.

In conclusion, the FF configuration used in a specific simulation, such as GAFF for small ligands and a protein-specific FF for proteins, is defined within the system’s topology file. This essential information remains constant throughout the simulation run. In the presented thesis, all the analyzed proteins were modeled using the AMBER99SB force field [71], an improved version of the original AMBER99 (released in 1999) that includes revised backbone torsion parameters for enhanced accuracy in protein simulations. Moreover, ligand parametrization was primarily performed using the GAFF force field with AM1-BCC charges. However, for the RNA-ligand systems discussed in Section 6.4.2, RESP charges and refined parameters were additionally employed.

Atomic point charges

By definition, a molecule consists of atoms whose electrons and nuclei generate an electrostatic potential at every point within the molecular space. For a positive test charge, electrons contribute with a negative, attractive interaction while the

converse holds for nuclei. Because the electronic density is continuous and cannot be easily represented by a simple analytical function, interactions are typically modeled using point charges centered at atomic positions. These point charges are chosen to best fit the QM electrostatic potential (ψ_{esp}), which comprises electronic ($\psi_{electrons}$) and nuclear (ψ_{nuclei}) contributions. Additional charges can be also incorporated to refine the model. The electrostatic potential, ψ_{esp} , felt by a unitary positive test charge at position r is given by the following equation:

$$\psi_{esp}(r) = \psi_{electrons}(r) + \psi_{nuclei}(r) = \sum_i^N \frac{Z_i}{|R_i - r|} - \int \frac{\rho(r')}{|r' - r|} dr' \quad (2.35)$$

Here, the positive contribution from the nuclei (ψ_{nuclei}) is calculated as the sum of the charges (Z) of all nuclei (N) divided by the distance between their position vector R_i and the position vector r where ψ_{esp} is evaluated. In contrast, the negative contribution from the electrons ($\psi_{electrons}$) is represented as an integral, reflecting their non-fixed positions as they exist in a cloud-like distribution around the nucleus. Here, $\rho(r')$ denotes the electron density, with r' being the position vector of the electrons.

The electrostatic potential ψ_{esp} is evaluated outside a surface around the molecule. This surface is often the Connolly surface [77], which provides an analytical, solvent excluded, representation of the molecular surface. The molecule is immersed in a 3D grid of points, and ψ_{esp} is calculated at each point of this grid but outside the surface. Point charges are determined to accurately reproduce ψ_{esp} at each grid node outside the surface, thus modeling the interaction between two specific point charges using Coulomb's law. The goal is to reproduce this potential by assigning fixed partial charges to the atoms of the molecule.

Thus, the objective becomes finding a set of charges that best matches the electrostatic potential calculated from QM. To compute the point charges (starting from the initial estimate), an error function $ErrF$ is defined. It quantifies the difference between the calculated and desired electrostatic potentials, and then this quantity is minimized. This procedure optimizes the selected charges ensuring they closely

match the reference ψ_{esp} . The error function formula is:

$$ErrF(Q) = \sum_j^{points} \left(\psi_{esp}(r_j) - \sum_i^N \frac{Q(R_i)}{|R_i - r_j|} \right)^2 \quad (2.36)$$

where the subscript *points* refers to the total number of points outside (or on depending on the fitting procedure) the molecular surface. The first term of Eq. 2.36 within the parentheses represents the reference QM electrostatic potential, while the second term corresponds to the electrostatic potential derived from point-charge model. By minimizing this error function, the sum of the squared differences between the two potentials is reduced.

However, it has been observed that atoms close to the surface significantly influence the potential values, whereas buried atoms have a minimal impact. This disparity can lead to inaccuracies and unrealistic charge values. Also the problem is a classical inverse problem hence its solution is unstable upon little changes of the potential. Many different charge distributions (solutions) can lead to the same potential; this requires defining an a-priori preferential solution among all the possible ones. To mitigate this issue, various restraints and constraints can be implemented:

1. Restrained Electrostatic Potential (RESP) methodology [75]: this method involves a quadratic regularization term (a Gaussian prior) to prevent atomic charges deviate too far from zero. This approach effectively minimizes unrealistic charge values, particularly for atoms that are deeply buried within the molecular structure.
2. Symmetry constraints involve applying constraints that enforce the equality of charges on atoms arranged symmetrically. This method ensures that the charges remain consistent across symmetric positions.
3. The total charge (i.e. sum of all point charges) is constrained to match the total net molecular charge.

Even if the point charge model is effective, it has some limitations. Indeed, it primarily focuses on replicating intermolecular interactions, which can result in unreliable predictions for intramolecular forces. Additionally, it completely neglects

the charge-conformational coupling, and atomic charges are not fully transferable.

In the end, it is important to emphasize that MD simulations generate molecular ensembles shaped by the underlying theoretical model, namely the chosen FF. The quality of this model directly determines the accuracy of the predicted statistical ensemble. Therefore, the reliability of the simulation results depends heavily on the quality of the FF parameters and the charge method selected. The process of parameterizing the FF, which involves determining all the necessary parameters for the FF equations, is time-consuming and requires substantial QM or experimental data, which may not always be readily available.

2.3.3 System preparation

To perform a MD simulation, the first essential requirement is an initial molecular conformation of the system of interest. This structural starting point is typically obtained from experimental methods such as X-ray crystallography, NMR spectroscopy, or cryo-electron microscopy. However, raw experimental structures are often not immediately suitable for simulation and must undergo extensive pre-processing [22]. These structures are usually provided in the Protein Data Bank (PDB) format, containing atomic Cartesian coordinates [78]. Before initiating a MD simulation, critical preparation steps are necessary to ensure the structural integrity and completeness of the molecular model [79]. These steps commonly include the removal of crystallographic artifacts such as crystallization co-factors and irrelevant ligands, the reconstruction of missing atoms, side chains, or flexible loops, and the capping of protein termini with neutral (uncharged) groups to avoid unwanted zwitterionic effects. In case water molecules are found these often kept if placed nearby the sites of interest (e.g. binding site). One of the most important yet missing components in experimental structures is hydrogen atom. Since many structural determination techniques do not resolve hydrogen positions, they must be added in-silico. Hydrogens are essential for accurate FF parameterization and are crucial in forming hydrogen bonds, which significantly contribute to the overall stability of biomolecular systems. Moreover, they play a key role in interesting phenomena such as protein folding and receptor-ligand binding.

The proper preparation of the initial structure ensures the reliability and physical accuracy of the subsequent MD simulations, ultimately leading to more meaningful insights into the molecular behavior of the system [57, 80, 79, 22].

Protonation state

A key step of system preparation for MD simulations involves determining the protonation states of protein’s amino-acids and, if applicable, the tautomers of histidine residues. The protonation of titratable amino acids such as Arginine (Arg), Lysine (Lys), Tyrosine (Tyr), Cysteine (Cys), Histidine (His), Glutamic acid (Glu), and Aspartic acid (Asp) is influenced by the surrounding micro-environment and pH. In highly polar environments, the charged state is more stable, whereas neutral forms are favored in less polar surroundings. At physiological pH (pH close to 7), Tyr, Lys, Cys, and Arg typically retain their standard protonation states, whereas Glu, Asp, and His may adopt non-standard forms depending on the context.

The protonation states of proteins’ residues plays a critical role in both their catalytic activity and structural stability. In standard MD simulations, correctly determining protonation states is a prerequisite for obtaining reliable results. Incorrect protonation states can significantly alter the protein behavior, potentially leading to invalid findings. Many studies have highlighted the critical role of protonation state assignment in MD simulations, emphasizing its necessity for achieving realistic and accurate results [81, 82].

Standard MD typically assume fixed protonation states throughout the simulation run. The most accurate approach to address this limitation involves constant pH simulations, where ionization states are dynamically adjusted during the simulation. However, this method is highly computationally demanding and not commonly employed. For routine applications, computationally efficient tools are available to estimate the pKa values of protein residues in the starting structure. These include both web-based platforms and standalone software, which generally employ one of two main approaches: empirical methods, such as PROPKA, or Poisson-Boltzmann-based calculations. It is important to note, however, that no

pKa prediction method is perfect. Using fixed protonation states is efficient but do not consider changes in protonation pattern that often accompany conformational rearrangements in biological systems. This limitation can hinder the understanding of proton-coupled dynamics. For studies where such dynamics are critical, constant pH simulations are a better alternative [83].

In the present study, the protonation states of residues were assigned using the GROMACS pdb2gmX program [84, 79] and the Protein Preparation Wizard of Schrödinger’s Maestro software. Specifically, Lys was protonated with three protons, Asp and Glu were left unprotonated, and Hist protonation was determined based on the optimal hydrogen bonding configuration. The pdb2gmX program automatically selects the charged forms of Lys, Asp, and Glu, while for His, it places the proton on either ND1 or NE2 depending on the hydrogen bonding network. Additionally, the N- and C-termini were ionized as NH_3^+ and COO^- , respectively [79].

2.3.4 Thermostats and Barostats

Realistic physical or chemical experiments are typically conducted at constant temperature, pressure, or volume. To closely replicate these conditions in MD simulations these parameters must be specified. Thus, by simulating the appropriate statistical ensembles, simulation results can be compared with experimental data [79]. Particularly, controlling T is crucial because many experimental observables, which can be directly cross-referenced to MD data, are highly sensitive to this parameter [85].

The direct integration of Newton’s equations samples the microcanonical ensemble, but precise prescriptions allow extensions to the other thermodynamic ensembles [46]. The two most widely used ensembles for MD simulations are the NVT ensemble, (canonical or isochoric-isothermal ensemble), and the NpT ensemble (isobaric-isothermal ensemble). To maintain a constant T in the system throughout the simulation, sophisticated algorithms known as thermostats are employed. Similarly, controlling the P involves using algorithms called barostats. In simple terms, thermostats and barostats act as virtual environments that ensure macroscopic T

and P to remain constant during the simulation time.

In MD simulations, a key energy component is the kinetic energy of the system, which is determined by the momenta of the particles:

$$K = \frac{1}{2} \sum_i m_i v_i^2 \quad (2.37)$$

In MM, the macroscopic T of a given system is linked to the time average of its kinetic energy $\langle K \rangle$, which in turn is related to the atoms' velocities. For a system composed by N particles, the equipartition theorem states that the average kinetic energy is expressed as:

$$\langle K \rangle = \frac{3}{2} N k_B T \quad (2.38)$$

To keep T approximately constant, the average kinetic energy must be fixed, while the instantaneous kinetic energy is allowed to fluctuate. Thus, one of the most naive approaches to regulate T is the velocity rescaling method, which represents the basic thermostat algorithm. This mathematical tool, called Berendsen thermostat (see later), is able to adjust the velocities in a way that produces a constant average kinetic energy over a certain period of time, ensuring T remains stable during the simulation. This condition is obtained by multiplying the velocities by a scaling factor at each simulation step. Even if this procedure is the basic idea behind thermostat algorithms, this simple method is not able to reproduce the correct canonical ensemble. Indeed, velocities are rescaled at each step, avoiding fluctuations in the kinetic energy. This behavior is not the desired one, making simple velocity rescaling algorithm rarely used in MD experiments. Like the simple velocity rescaling method, all thermostats introduce perturbations to the equations of motions that affect both the dynamical and equilibrium properties of the system. Moreover, the T distribution must be homogeneous within the simulated solute-solvent system. Often problems arise in controlling temperature for inhomogeneous systems, as different regions may experience varying degrees of heating, leading to problematic scenarios such as the "hot solvent and cold solute" situation [85]. A common strategy to address this issue is to couple the different sub-parts of the system to two separate thermostats, allowing for independent T

control.

Likewise, barostats are mathematical algorithms designed to allow instantaneous P fluctuations while maintaining a constant average pressure over time. Instead of adjusting particle velocities, as thermostats do, barostats scale the volume of the simulation box. When the box size decreases, atoms become compressed, resulting in an increase in the overall P . Conversely, expanding the box size enables atoms to occupy a larger volume, which reduces P . In practice, this volume adjustment is achieved by scaling the atomic coordinates of each atom according to the desired P adjustments.

Different types of thermostats and barostats are available for MD simulations, each offering unique advantages and disadvantages. Below, an overview of the most commonly used algorithms in MD simulations is reported.

Berendsen thermostat

The Berendsen Thermostat [86] is a popular algorithm for T control, achieved by rescaling particle velocities. It mimics a weak coupling with first-order kinetics to an external heat bath at a desired temperature T_0 . The T correction is expressed by the following equation:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} \quad (2.39)$$

where τ is the time constant that controls the coupling strength. T deviation decays exponentially with the time constant τ . The algorithm adjusts the system's kinetic energy at each time step by modifying particle velocities accordingly, and it is integrated numerically at every simulation step. Although widely used, the Berendsen thermostat introduces some artifacts and does not sample the canonical ensemble accurately. Specifically, it suppresses kinetic energy fluctuations and violates energy equipartition, which can lead to incorrect sampling. The extent of this issue decreases with the system size, as the error scales with $1/N$, but fluctuation properties remain affected. To address these problems, Bussi *et al.* [87] proposed a modified version of the Berendsen scheme that provides better ensemble sampling.

Bussi-Donadio-Parrinello thermostat

The Bussi-Parrinello thermostat [87] builds on the Berendsen thermostat by introducing an additional stochastic term that ensures a correct kinetic energy distribution. This modification guarantees that the system samples from a proper canonical ensemble, overcoming the limitations of the original Berendsen approach. This thermostat does not require additional parameters except for a random seed. It retains the benefits of the Berendsen thermostat, such as first-order decay of T deviations and the absence of oscillations, while ensuring accurate thermodynamic sampling.

Bussi-Parrinello algorithm was used to control the T in all the MD simulations performed in the presented study.

Berendsen barostat

The Berendsen barostat [86] is conceptually similar to the corresponding thermostat. The system is treated as weakly coupled to a pressure bath, where the volume is adjusted by applying a scaling factor. The algorithm modulates the simulation box size in response to P changes, ensuring that the system maintains the average P . However, this approach has some disadvantages. Particularly, it does not truly sample from the real NPT ensemble. As a result, it is commonly employed during the equilibration phase, as was done in our study, but it is not suitable for the production phase of MD simulations.

2.3.5 Periodic Boundary Conditions (PBC)

In molecular simulations, boundaries are fundamental to maintain thermodynamic properties like T and P . However, due to the small system size in common MD set-up, a molecular cluster often has substantial unwanted boundaries with the surrounding vacuum. To alleviate this problem, periodic boundary conditions (PBC) have been introduced, enabling to simulate the bulk solvent (or crystalline environments) within the simulation. Particularly, PBC approximate an infinite system by replicating a smaller finite section of the system (known as the unit cell or the

periodic box) in the space. This arrangement allows atoms within the periodic box to be surrounded by identical, translated copies of itself, minimizing edge effects and accurately representing bulk properties [57, 22]. With PBC, the simulation box (with all its content) is replicated in all directions, creating an infinite lattice where particles' movements in one box are mirrored in all others. The particle count in each unit cell is maintained constant. This approach eliminates unnatural boundary artifacts, though the periodicity can introduce minor errors, generally smaller than those caused by simulating an isolated system with edge effects [88, 22].

Popular simulation engines, like GROMACS [84], implement PBC joined with the minimum image convention. This approach ensures that only the nearest periodic image of each particle is considered for short-range non-bonded interactions, such as dipole-dipole and van der Waals forces. Thus, only one image of each molecule in the PBC is considered for a pair interaction. Each particle only interacts with atoms within a box centered on itself and matching the original box's size and shape, limiting interactions to the closest periodic images rather than calculating an infinite interaction set.

In contrast, long-range electrostatic interactions (e.g., interactions involving net charges) are computed through lattice sum methods, such as the Ewald summation [89] or, more commonly, the Particle Mesh Ewald (PME) method [90]. PME is widely favored as it separates charge interactions into two terms compatible with PBC, simplifying calculations by reducing the impact of long-range interactions into manageable components. Particularly, the Coulomb potential is split into two terms: a short-range term, calculated within a defined cutoff radius in real space, and a long-range term, evaluated in reciprocal space using Fourier transforms. This approach enables an efficient treatment of electrostatic interactions under periodic boundary conditions. [90, 22].

2.3.6 Water models

Water plays a central, ubiquitous role in biological phenomena, making the accurate representation of aqueous environments essential for reliable simulations of biological processes, particularly in protein-ligand binding [91, 92]. Since water

typically constitutes over 90% of a solvated system, it contributes significantly to the potential and kinetic energy of the system. Consequently, its correct modeling is crucial for obtaining quantitative results from simulations [93, 94].

Fully atomistic MD requires well-developed FFs for biomolecules and solvent models, with various theoretical approaches available to describe the aqueous environment. Empirical models used to represent aqueous effects can be categorized into three main families: implicit water models or continuous solvents, explicit water models, and coarse-grained water models. In this study, we focus solely on the second category of solvent models, while a detailed discussion of implicit solvents is available in the literature [95].

Explicit water models use discrete solvent molecules, offering the highest level of detail for atomistic simulations. Although more computationally demanding than implicit models, they provide greater accuracy in describing solvation effects [11]. These empirical models are designed to replicate various bulk properties of water in a specific phase. However, none of these models can perfectly replicate all experimental properties of bulk water; i.e., Static dielectric constant, Self diffusion coefficient, Heat of vaporization, Isobaric heat capacity, Thermal expansion coefficient, Isothermal compressibility [96]. Consequently, choosing the right model requires ensuring it accurately represents the relevant properties of water for the studied phenomena [93, 94].

Water monomers have a simple chemical structure consisting of two hydrogen atoms and one oxygen atom, with an OH bond length of 0.9572 Å and an HOH angle of 104.52°. Early water models employ a rigid geometry with harmonic constraints resembling this real geometry of water molecules. Meanwhile, the electron density distribution is represented by point charges to reproduce the electrostatic potential of the molecule. The current section provides a summary of the key features of the most recent and commonly used explicit water models for protein MD simulations.

Explicit water models consist of a FF potential energy function and parameters that define intra- and intermolecular interactions in water clusters and aqueous environments. The most utilized category of explicit water models is the Fixed-Charge

N-point Water Models family, where N is an integer ranging from 3 to 6 indicating the number of interaction sites. Their FF equation involves only non-bonded terms. Particularly, the Coulombic potential for pairwise interactions between intermolecular pairs of point charges, and the LJ 12-6 potential for dispersion and repulsion interactions. The general pairwise potential energy function for these models is defined in the following equation:

$$V = \sum_{pairs} \left(\frac{A_{LJ}}{r_{OO}^{12}} - \frac{B_{LJ}}{r_{OO}^6} + k \frac{q_i q_j}{r_{ij}} \right) \quad (2.40)$$

where r_{OO} is the oxygen-oxygen distances, A_{LJ} and B_{LJ} are the LJ parameters, q_i and q_j are the charges of the i^{th} and j^{th} charged sites at distance r_{ij} , and k is the electrostatic constant of the Coulombic potential. The geometric and energy function parameters are fitted to match specific experimental properties of water, usually at the expense of others. As can be seen, Eq. 2.40 does not explicitly include many-body or QM effects. Additionally, it lacks terms for considering geometry changes as bond stretching or angle bending, keeping the water molecule's geometry fixed even after interactions (such as dimer formation). As a result, these missing physical effects are approximated using a mean-field averaging approach. Building on these original rigid models, flexible variants have been developed to account for water molecule geometry fluctuations using harmonic potentials.

Over the years, various models have been developed, each with parameters tuned to replicate specific experimental properties of water. However, the parameters of a specific water model are often sub-optimal in reproducing jointly all experimental characteristics of water, leading to inaccuracies in MD results [97]. Although more advanced and polarizable water models have been developed, fixed-charge 3- and 4-point models remain the most widely used in protein-ligand binding simulations, as they provide a balance between accuracy and computational efficiency. Their most important properties are reported in Table 2.1.

3-point water models

Water molecules in three-point models are modeled by using a single oxygen and two hydrogen atoms, each assigned a fixed partial charge at the atomic nuclei. These models are characterized by six independent parameters: the OH bond length, the HOH bond angle, partial charges on oxygen and hydrogen atoms, and two LJ parameters. The LJ interaction site is assigned to the oxygen atom, while hydrogen atoms lack explicit van der Waals interactions. By following different parameterization strategies, several 3-point models have been developed. Table 2.2 summarizes the FF parameters for the commonly used three-point water models.

One of the earliest and most widely adopted explicit solvent models in biomolecular simulations is the Transferable intermolecular Potential 3-Point (TIP3P) model. Developed based on experimental gas-phase water geometry [98], TIP3P is a rigid, non-polarizable model that accurately reproduces key bulk water properties, such as the enthalpy of vaporization and the dielectric constant at 298 K and 1 bar. Nevertheless, it tends to underestimate density while overestimating the self-diffusion coefficient. Despite these limitations, TIP3P remains a popular choice for simulating solvated biomolecules due to its simplicity and computational efficiency. Therefore, our systematic investigation of water models, reported in Chapter 5 [P2], began with simulations conducted in a simple TIP3P environment.

The flexible variant of TIP3P, known as TIP3P/Fw, was later introduced to incorporate intramolecular flexibility by adding additional potential energy terms [99]. While this modification improves the representation of water interactions, it also increases computational cost. Specifically, the introduction of flexibility requires reducing the integration time step from 2 fs, typically used in rigid models with bond constraints, to 1 fs, effectively doubling computational expense. However, this added flexibility does not always result in a substantial improvement in the accuracy of water behavior modeling [P2].

A more recent advancement in three-point water models is the Optimal Point Charge (OPC3) model [100]. Unlike TIP3P, which relies on predefined charge positions and geometric constraints, OPC3 is designed to more accurately represent

the electrostatic field, leading to improved modeling of hydrogen bonding and other essential liquid water properties. As highlighted in the study of Onufriev *et al.* [97], OPC3 outperforms many widely used models by providing superior accuracy across a broad temperature range. By eliminating fixed point charge geometries, OPC3 represents the most precise non-polarizable three-point model currently available for describing bulk water properties [100].

While three-point water models offer a balance between computational efficiency and accuracy, they have some limitations in reproducing specific characteristics of liquid water. To address these shortcomings, four-point models have been developed.

4-point water models

Water monomers in four-point models are represented by one oxygen atom, two hydrogen atoms, and an additional massless dummy atom. Unlike three-point models, where the negative charge is assigned to the oxygen atom, four-point models reposition this charge slightly (by 0.15 Å) along the bisector of the HOH angle. This adjustment, represented by the dummy atom, enhances the model’s ability to reproduce the quadrupole moment of water, thereby improving the accuracy of its electrostatic potential. Four-point water models are characterized by seven parameters: the OH bond length, the HOH bond angle, the oxygen-dummy atom distance, partial charges on the hydrogen and dummy atoms, and two LJ parameters. Various parameterization strategies have resulted in multiple variants, such as TIP4P, its flexible version TIP4P/Fw, and OPC. The force field parameters for these models are summarized in Table 2.3.

The Transferable intermolecular Potential 4-Point (TIP4P) model was developed as an improvement over TIP3P. While maintaining a rigid three-charge planar structure, TIP4P introduces a dummy atom to carry the negative charge, which enhances the water dimer geometry and the representation of the electrostatic distribution [101]. However, this modification results in an underestimated dielectric constant.

A flexible version, TIP4P/Fw, was later introduced, incorporating a harmonic

term for angle bending and a Morse potential for bond stretching [102]. In addition, minor adjustments were made to the Lennard-Jones parameters to improve intermolecular interactions. However, like other flexible water models, TIP4P/Fw increases computational cost without consistently yielding better accuracy [103] [P2].

Finally, the Optimal Point Charge (OPC) model represents a departure from traditional geometric constraints [104]. It served as the foundation for the subsequent OPC3 model, with its parameters optimized to more accurately reproduce the electrostatic potential of water molecules, while eliminating fixed charge placement restrictions. This approach leads to significantly improved accuracy compared to other models [104].

Beyond four-point models, water FFs with additional interaction sites have also been developed to improve the representation of certain bulk properties. For example, the Transferable intermolecular Potential 5-Point (TIP5P) model incorporates the non-planar charge distribution of water. However, adding more interaction sites increases computational costs and parameterization complexity, with unclear benefits for biomolecular simulations. Additionally, numerous other water models have been designed to optimize particular properties or to incorporate polarization effects. A more detailed discussion of these advancements can be found in literature [96].

All the presented water models are widely used to represent aqueous solvation effects in MD simulations. Nevertheless, since the accuracy of atomistic MD simulations relies on the precise treatment of solvation effects, it is essential to assess the robustness of simulation outcomes concerning the chosen water model. Moreover, it is important to note that water models have traditionally been parameterized for room temperature (298 K) and without accounting dissolved solutes. Indeed, ensuring consistency between the chosen water model and the protein FF is also crucial for reliable simulation results. For example, TIP3P has been widely used with AMBER FFs, which have been specifically optimized for compatibility with this water model. The AMBER team has recently recommended using OPC water models in conjunction with their latest biomolecular FFs. In contrast, according to

MODEL	Density (g/cm ³)	Coord. n°	Surface tension (mN/m)	Dielectric constant	Self- diffusion coeff. (*10 ⁻⁵ cm ² /s)
TIP3P	0.980 ± 0.006	6.239 ± 0.001	47.0 ± 0.2	95 ± 3	5.72 ± 0.04
TIP3P/Fw	1.027 ± 0.006	5.3 ± 0.3	55.2 ± 0.4	197 ± 2	3.8 ± 0.1
OPC3	0.991 ± 0.006	4.9 ± 0.2	61.0 ± 0.2	79 ± 1	2.28 ± 0.02
TIP4P	0.994 ± 0.006	5.14 ± 0.001	52.2 ± 0.2	51.3 ± 0.5	2.57 ± 0.04
OPC	0.997 ± 0.005	5.1971 ± 0.0005	70.1 ± 0.2	78 ± 1	2.27 ± 0.02
Exp	0.997	4.7	71.99	78.3	2.30

Table 2.1: Key properties of the most commonly used water models, with corresponding experimental values listed in the final row.

MODEL	O σ (Å)	O ϵ (kcal/- mol)	O charge (e)	H charge (e)	OH bond (Å)	H-O-H angle (°)
TIP3P	3.15061	0.1521	-0.834	0.417	0.9572	104.52
TIP3P/Fw	3.15061	0.1521	-0.834	0.417	0.9572	104.52
OPC3	3.17427	0.163406	-0.89517	0.447585	0.97888	109.47

Table 2.2: Force field parameters for 3-point water models. For the flexible model, the additional parameters are $k_b = 529.581$ kcal mol⁻¹Å⁻² and $k_\theta = 50.0$ kcal mol⁻¹rad⁻².

MODEL	O σ (Å)	O ϵ (kcal/- mol)	O charge (e)	H charge (e)	OH bond (Å)	H-O- H angle (°)	OM bond (Å)
TIP4P	3.1565	0.155	-1.04	0.52	0.9572	104.52	0.15
TIP4P/Fw	3.1565	0.155	-1.04	0.52	0.9572	104.52	0.15
OPC	3.16655	0.212801	-1.3582	0.6791	0.8724	103.6	0.1594

Table 2.3: Force field parameters for 4-point water models. For the flexible model, the additional parameters are $k_b = 103.389 \text{ kcal mol}^{-1} \text{Å}^{-2}$ and $k_\theta = 2.287 \text{ kcal mol}^{-1} \text{rad}^{-2}$.

GROMACS recommendations [84], TIP4P results more suitable when paired with the OPLS FF for proteins. Nevertheless, the optimal combination of protein and water FFs for specific applications is often not well-established.

Finally, a good water model strikes a balance between accuracy and computational efficiency, as interactions among water molecules significantly impact the overall simulation cost. The computational expense is closely tied to the number of pairwise distance calculations required within the model. In a 3-point model with three charge sites, nine pairwise distances must be computed per water molecule pair. In contrast, 4-point models require ten distances, including three Coulomb interactions and one van der Waals O-O interaction. This underscores the trade-off between model complexity and computational demand, highlighting the importance of selecting a water model suited to the specific needs of the simulation.

2.3.7 MD output and conclusion

Lastly, the final output of a MD simulation is a trajectory file containing the positional coordinates and velocities of atoms recorded at regular intervals. This file effectively captures snapshots of the system throughout the simulation, producing a three-dimensional "movie" that describes the molecule's configuration over time. To manage file size, trajectories are typically saved every 10 to 15 steps, ensuring that each frame retains the essential data without overwhelming storage [22].

Well-designed MD simulations provide detailed structural insights into the biomolec-

ular system under investigation, enabling the analysis of conformational changes or protein-ligand binding events. However, interpreting these results can be complex and must account for the limitations inherent in the numerical methods. Despite significant advancements in MD algorithms and computational power, most biological simulations still only capture few microseconds of real-time, often requiring extensive computational resources over several days or weeks.

2.4 Rare events & Enhanced sampling

The dynamics of biomolecular systems occur over a broad spectrum of both time and spatial scales, reflecting the diversity of molecular motions involved. At the shortest time scale, femtosecond dynamics capture atomic vibrations. Picosecond to nanosecond events include bond rotations, small side-chain rearrangements, and local conformational changes in proteins. On the longer end, microsecond to millisecond (even second) timescales cover larger, more complex processes such as loop movements, protein folding, large-scale domain reorientation, and ligand binding. These slower processes, often referred to as rare events, are crucial to understanding biomolecular function and stability [4, 22].

Rare events occur on long timescales because of the high free energy barriers (greater than $k_B T$) that the system must overcome to transition between its stable states. This complex configuration landscape consists of hierarchically organized basins separated by high energy barriers, which can disrupt ergodicity [4, 105]. These rare events often extend beyond the timescales accessible to traditional MD simulations. While individual MD time steps are typically 1-2 femtoseconds, advancements in parallel computing and integration of graphical processing units (GPU) have enabled simulations to reach the microsecond scale. However, reaching millisecond to second timescales remains a significant challenge and is generally beyond the capabilities of standard research setups [105, 106]. An exceptional example is the work by the D. E. Shaw Research group, which developed a specialized MD-focused hardware system called Anton [107]. This high-performance, expensive machine has achieved simulations on the millisecond scale, effectively capturing

complex processes like protein-ligand binding, and underscoring the reliability and potential of MD in rational drug discovery.

To address timescale limitations, several so-called, *enhanced sampling* methods have been developed. They accelerate rare events and facilitate the exploration of complex configuration landscapes [108, 106]. Enhanced MD simulations follow two main routes: (i) alchemical transformations, and (ii) path-based/geometrical transformations. Moreover, they often relies on order parameters to capture key aspects of the system’s behavior.

2.4.1 Order parameter: coupling parameter & CV

In 1920, De Donder introduced the concept of the degree of advancement of a chemical process, a mathematical tool to describe the extent of a chemical reaction [109]. This concept determines the direction of the chemical reaction, with the degree of advancement represented as a variable along an axis from left to right [110]. Starting from this original idea, Kirkwood introduced the concept of order parameter which is a variable describing changes in the system’s state over time [111]. This mathematical approach provides insight into the progression and directionality of the studied chemical transformation [109]. When the order parameter corresponds to the pathway of the system evolution, it is called reaction coordinate. Therefore, the order parameter acts as an approximate reaction coordinate, and is conceptually connected to the committor function. The committor quantifies the probability that a given molecular configuration will reach the product state before returning to the reactant state. In many contexts, it can be seen as the optimal reaction coordinate providing the most accurate description of transition progress. As such, committor analysis is often used as a benchmark for assessing the quality of approximate reaction coordinates.

In the context of free energy calculations, the order parameter is selected to characterizes and capture the essential changes occurring within the system during a transformation. Its specific role and interpretation differ depending on the methodological approach: in (i) alchemical transformations, it typically represents a non-physical transition between molecular states; whereas in (ii) path-based or geo-

metrical methods, it corresponds to a physical coordinate or reaction path along which the system evolves.

When alchemical transformations are performed, the order parameter is called coupling parameter and often represented by the λ symbol $\in [0, 1]$. This parameter supports the interpolation between two states (e.g., ligand bound vs. unbound) by gradually scaling interactions from one state to the other. In other words, the coupling parameter describes the interpolation between the Hamiltonians of the initial and final states.

In contrast, in path-based approaches the order parameter defines the collective variables (CV)s. CVs (also denoted by ξ) represent physical coordinates tracking the progress along a reaction pathway. The variation of these CVs provides a quantitative measure of the system’s progression along the reaction pathway under investigation. Formally, a CV is a mathematical function that projects the high-dimensional configurational space of the system (with N degrees of freedom) onto a lower-dimensional space, making complex dynamics more tractable [108]. CVs are chosen based on the relevant degrees of freedom of the system associated with the process of interest. They can include simple metrics such as interatomic distances, or root-mean-square deviation (RMSD), as well as more complex observables. In difficult transformations, such as protein-ligand binding, simple geometric descriptors like distances or angles often fall short in representing the true mechanistic evolution of the process. In such cases, more sophisticated CVs are required to account for the relevant degrees of freedom and the collective behavior of the system. Once a suitable CV is identified, it is treated as an external parameter and incorporated into the Lagrangian of the system, typically through the addition of a biasing potential.

The result of path-based calculations is typically the free energy profile along the chosen CVs, which is named the potential of mean force (PMF)[108, 112]. PMF represents the Landau free energy profile along the CV which is used in the enhanced sampling procedure. The concept of the PMF (first introduced by Kirkwood) has its origins in the study of simple liquids, where it was historically used to describe their structural properties. Moreover, while the free energy is formally defined as

the logarithm of the full partition function (Eq. 2.10), the Free Energy Surface (FES) corresponds to the logarithm of a partially integrated partition function, where the integration is performed over all degrees of freedom except the selected CVs [108]. The two terms, PMF and FES, are often used as synonyms in the literature [113]

Therefore, to effectively use MD simulations in drug discovery, enhanced sampling methods and effective order parameters are needed. As mentioned before, the protein-ligand binding event takes long time to occur due to the low probability of crossing high free energy barriers separating low energy regions. To address this, dedicated enhanced sampling algorithms and advanced order parameters have been developed. These methods can capture binding and unbinding events, unlike traditional MD simulations [114].

Path Collective Variables

For particularly challenging transformations, such as protein-ligand binding, designing effective CVs is essential. Simple metrics may not accurately capture the complex evolution of a system, thus more sophisticated CVs may be needed to represent the relevant degrees of freedom. For example, as a first approximation, the distance between two molecules can be a straightforward CV in a protein-ligand binding analysis. However, more intricate CVs as Path Collective Variables (PCVs) [115] and CVs derived from dimensionality reduction techniques can better describe the collective dynamics.

PCVs have been developed by Branduardi *et al.* [115], and they describe the position of a point in the configurational space relative to a predefined path. In detail, the PCVs are two: $S(x)$, measuring the progression along the high-dimensional path, and $Z(x)$, measuring the orthogonal deviation with respect to it:

$$S(x) = \frac{\sum_{i=1}^p i e^{-\lambda \|x-x_i\|^2}}{\sum_{i=1}^p e^{-\lambda \|x-x_i\|^2}} \quad (2.41)$$

$$Z(x) = -\lambda^{-1} \ln \sum_{i=1}^p e^{-\lambda \|x - x_i\|^2} \quad (2.42)$$

Here, λ is a smoothness parameter, p is the number of molecular configurations included in the predefined path, and $\|x - x_i\|^2$ quantifies the distance between the i^{th} structure in the path (x_i) and the instantaneous microscopic configuration (x). Different distance metrics can be used to define this distance [116], as the RMSD or the Mean-square deviation (MSD). Note that when $x = x_i$, $S(x_i)$ acquires approximately the integer value i , rendering S an indicator function of the configuration progression. $S(x)$ acts as a faithful representation of the physical configurations, provided it is smooth and creates a linear mapping from the configurational space to the progressive values of S . To achieve this accurate mapping, one can parameterize λ using the following equation:

$$\lambda = \frac{2.3 p}{\sum_{i=1}^{p-1} \|x_i - x_{i+1}\|^2} \quad (2.43)$$

where $\|x_i - x_{i+1}\|^2$ quantifies the distance between the i^{th} and the $i+1^{th}$ molecular structure. This equation was derived heuristically by the authors of the PCVs [117]. Finally, a uniform spacing between consecutive configurations in the predefined path is a necessary condition to apply PCVs correctly.

2.4.2 Enhanced sampling methods

Several approaches can be used to accelerate sampling, including adding an external potential to the system, applying targeted restraints to gather data at specific points in phase space, or gradually adjusting the system’s Hamiltonian to match a reference state. These techniques not only speed up sampling but, in some cases, also enable the reconstruction of free energy landscapes [108].

To mitigate the inefficiencies of standard MD, in 1977 Torrie and Valleau introduced the Umbrella Sampling (US) method. [118]. US is historically the first algorithm developed to enhance the sampling of rare events, facilitating the exploration of high-energy or low-probability states within molecular systems. In

standard simulations, the regular potential $U(\mathbf{r})$ leads to a Boltzmann-weighted sampling. Conversely, US applies a simple harmonic restraint with strength k centered at ξ_0 :

$$\hat{U}(\mathbf{r}) = U(\mathbf{r}) + \frac{1}{2}k(\xi(\mathbf{r}) - \xi_0)^2 \quad (2.44)$$

The modified potential $\hat{U}(\mathbf{r})$ generates a sampling that enhances exploration of low probability regions of the configurational space, especially relevant for free energy calculations. The biasing is typically applied along a chosen reaction coordinate (ξ_0) or CVs to overcome energy barriers and improve the sampling of rare events. Thus, the selected CVs are restrained and pulled by the bias potential toward a target value. Since multiple staged simulations are performed with biased harmonic restraints centered at different ξ_0 values, US is considered a stratification strategy. This approach ensures sampling along the defined reaction coordinate. Finally, to recover the unbiased Boltzmann probability distributions and reconstruct the free energy profile along the CVs, various estimators have been devised [119, 120]. Starting from this original idea, different enhanced sampling methods have been developed to accelerate conformational sampling by various means.

Metadynamics

Among the most commonly employed enhanced sampling algorithms for protein-ligand binding simulations there are Metadynamics (MetaD) and metadynamics-based approaches [114]. MetaD was introduced by Laio and Parrinello in 2002 [121], and it was developed to help systems escape local free energy minima, enabling the exploration of free energy landscapes associated with rare events. In MetaD simulations, the system’s dynamics is biased along one or more CVs using a history-dependent bias potential. At set intervals, small Gaussian potentials centered at CV values, characterized by a width σ_i and height W , are added to the current energy values. This prevents the system from repeatedly visiting the same regions of configuration space. Gradually, this bias accumulates in the free energy minima, filling the underlying free energy profile and allowing the system to achieve a diffusive behavior along the CVs [122].

The history-dependent bias potential $U_G(\xi, t)$ at time t is defined by the formula:

$$U_G(\xi, t) = \int_t^0 \omega \exp \left(- \sum_{i=1}^d \frac{(\xi_i(t) - \xi_i(\tau))^2}{2\sigma_i^2} \right) d\tau \quad (2.45)$$

Here, ξ is the chosen CV, d is the dimension of the CV and ω is an energy rate calculated as the ratio of the Gaussian height W and the deposition time τ_G . The parameters needed for MetaD simulations, which are Gaussian height, width and deposition time, must be carefully chosen. At convergence, the bias potential serves as an unbiased estimator of the free energy surface, as described by the following relationship (where C is an arbitrary constant) [123]:

$$U_G(\xi, t \rightarrow \infty) = -F(\xi) + C \quad (2.46)$$

However, the original implementation of MetaD faced some limitations: determining when to stop a simulation is challenging, as the bias deposition is continuous, and there is no guarantee of convergence. To address these issues and improve the convergence of the free energy profile, Well-Tempered MetaD (WT-MetaD) was introduced [124, 125]. In WT-MetaD, the height of the Gaussian is decreased with simulation time. Other MetaD variations have also been developed, such as Multiple Walkers MetaD, which enhances parallel sampling [126], and the use of adaptive Gaussians that adjust dynamically to improve accuracy and convergence [127]. Moreover, that MetaD-related methods (even if they are enhanced sampling techniques) do not require too intense additional post-processing steps. They efficiently provide an unbiased estimate of the free energy directly projected onto the predefined CVs [114].

Since its introduction, WT-MetaD has been widely adopted in simulation; Gervasio *et al.* [128] were among the first to employ it for ligand-binding studies in 2005. A detailed review of MetaD-based algorithms used in drug discovery is available in reference [114]. Despite its success, MetaD-based methods present intrinsic limitations, most notably the challenge of identifying a suitable set of CVs capable of capturing the relevant slow degrees of freedom. In addition, the reliability of

the reconstructed PMF depends critically on allowing the system to equilibrate between successive additions of biasing potentials, a condition that is not always straightforward to satisfy in complex biological systems.

Steered Molecular Dynamics

In the late 1990s, Schulten’s group developed another enhanced sampling algorithm based on bias potentials, known as Steered Molecular Dynamics (SMD) [129, 130]. SMD is a nonequilibrium technique where a time-dependent harmonic potential ($U_\lambda(\xi, t)$) is summed to the system’s standard potential, similarly to US. However, in SMD, the center of the external potential shifts over time along the CV (ξ), effectively driving the system out of equilibrium. SMD is considered a nonequilibrium method because of its time-dependent harmonic potential:

$$U_\lambda(\xi, t) = \frac{1}{2}k(\xi - \xi_0(t))^2 \quad (2.47)$$

Originally, this approach was developed to explore rare biological events, such as protein folding [131] and protein-ligand binding [132]. Particularly, in the context of ligand unbinding, a spring with force constant k is virtually attached to the ligand, and the center of the restraint is pulled along the CV at a specific velocity v .

A few years later, in a foundational study, Park and Schulten [133] applied the Jarzynski relation for obtaining the free energy profile from nonequilibrium SMD simulations. Successively, in 2010, Colizzi *et al.* [134] pioneered the qualitative application of SMD in drug discovery to discriminate active from inactive compounds laying the groundwork for new structure-based drug design strategies utilizing SMD.

In the protein-ligand binding studies presented in this thesis, both WT-MetaD and SMD simulations were employed alongside PCVs to enhance the sampling of protein-ligand events.

Adiabatic Bias MD

While WT-MetaD and SMD are techniques that allow for efficient sampling of rare events and facilitate free energy calculations (using appropriate statistical estimators), methods such as adiabatic bias molecular dynamics (ABMD) [135] primarily serve an exploratory purpose. In ABMD simulations, a target value of the CV (ξ) is defined, and the protocol ensures that this value is reached within a finite time frame using a moving harmonic restraint [135]. The main benefit of this protocol is its *gentleness*. Indeed, the restraint center moves towards the target value only through thermal fluctuations, not through a prescribed time schedule as in SMD. The bias potential $U(t_n)$ at time t_n becomes:

$$V(t_n) = \alpha(\xi(t_n) - \xi_0)^2 \quad (2.48)$$

The update of the center ξ_0 is adjusted dynamically based on the progression of ξ in the intended direction. If an increase in ξ is wanted, the equations for updating the center are given by:

$$\xi_0 = \begin{cases} \xi_0 & \text{if } \xi(t_n) < \xi_0 \\ \xi_t & \text{if } \xi(t_n) \geq \xi_0 \end{cases} \quad (2.49)$$

From a pictorial perspective, ABMD can be compared to a pawl-and-ratchet mechanical system, where progress is incrementally secured, preventing backward motion. In this analogy, the CV is the wheel, which is allowed to move only in one direction. If the system attempts to move in the opposite direction, a harmonic restraint is applied to prevent this unwanted motion (with a user-defined restraint constant). As evident, this approach offers a gentler alternative to SMD. The key difference between ABMD and SMD is that in ABMD, the movement of the center is influenced not only by user-defined parameters but also by the system's natural progression toward the final CV value.

In this thesis, ABMD simulations were carried out using a custom-defined, electrostatics-inspired CV based on the Yukawa potential, for exploratory purposes.

2.5 Binding Free Energy Calculations

Free energy is one of the most fundamental thermodynamic quantities, typically represented as either Helmholtz or Gibbs free energy. A particular kind of free energy is binding free energy, which refers to the free energy difference associated with the binding of a potential therapeutic to a protein. This quantity directly reflects the drug’s affinity for its target, thus it serves as a critical parameter for optimizing new drug candidates [4].

Due to its relevance, accurately estimating binding free energy in protein-ligand interactions is essential for efficient *in-silico* drug discovery [11]. However, computing binding free energy from simulations pose many challenges. Particularly, its calculation requires sampling the full relevant configurational space of the system with good statistics, including crossing energy barriers between different states. Because of the complexity of the ligand-protein binding event, high energy regions may be under-sampled, while some low energy regions might be inaccessible due to trapping in local minima. This scenario makes accurately calculating binding free energy very difficult. Moreover, many existing free energy calculation methods face practical limitations in pharmaceutical applications, primarily due to convergence issues in the free energy estimation process.

As anticipated in section 2.1.1, the absolute Helmholtz free energy of a given system is:

$$F(N, V, T) = -k_B T \ln Q_{NVT} \quad (2.50)$$

where Q_{NVT} is the canonical partition function of the system defined as:

$$Q_{NVT} = \frac{1}{h^{3N} N!} \int_{r_i, p_i} e^{-\beta H(r_i, p_i)} dr_i^N dp_i^N \quad (2.51)$$

Eq.2.51 involves a $6N$ -dimensional integral, which is taken over all particle coordinates and their corresponding conjugate momenta. It is weighted by the Boltzmann factor of the system’s Hamiltonian, where the Hamiltonia represents the system’s total energy for a given configuration. Both Equations 2.50 and 2.51, previously

introduced in Section 2.1.1, are reiterated here for clarity.

Following equation 2.50, the direct calculation of absolute free energy is limited to simple cases where an analytical expression for Q_{NVT} can be derived. This is not the case for biological systems, where obtaining such analytical formulation is generally impractical. Absolute free energy can be calculated as the difference in free energy between the system of interest and a chosen reference state. Nevertheless, identifying an appropriate reference state for liquid-phase systems is often challenging. In the absence of a well-defined reference, the calculation of absolute free energies becomes impossible. However, in most situations, the objective is to compare results with experimental data, which leads researchers to focus on determining relative free energy differences. For biological systems, this could involve assessing the difference in binding affinities of two compounds to the same receptor. The binding free energy between the bound and unbound states of a given ligand to its target is defined as:

$$\Delta F_b = F_{bound} - F_{unbound} = -\frac{1}{\beta} \ln \frac{Q_{bound}}{Q_{unbound}} \quad (2.52)$$

where Q_{bound} and $Q_{unbound}$ are the partition functions associated to the bound and unbound states, respectively.

Most of the free energy methods used nowadays are built upon foundational principles established by early pioneers in the field, such as: John Kirkwood, Robert Zwanzig, Benjamin Widom, John Valleau, and Charles Bennett [105]. Moreover, computing binding free energy from computer simulations requires three basic components [136] (graphically summarized in Figure 2.2):

1. An appropriate **Hamiltonian** to model the system of interest.
2. A **sampling protocol** to generate a representative ensemble of configurations, ensuring exhaustive exploration of the conformational space.
3. A reliable **estimator** to calculate free energy differences, allowing accurate determination of the free energy from the sampled configurations.

In the context of MD simulations for binding free energy calculations, the first

ingredient (i.e., Hamiltonian to model the system) is a classical Hamiltonian. This classical Hamiltonian typically incorporates fixed point charges when modeling biomolecules, and is commonly referred to as a force field. As discussed in Section 2.3.2, different FFs may utilize different functional forms, but they fundamentally share the same underlying concepts. In practice, the choice of the Hamiltonian often requires a balance between accuracy and computational efficiency.

3 BASIC COMPONENTS

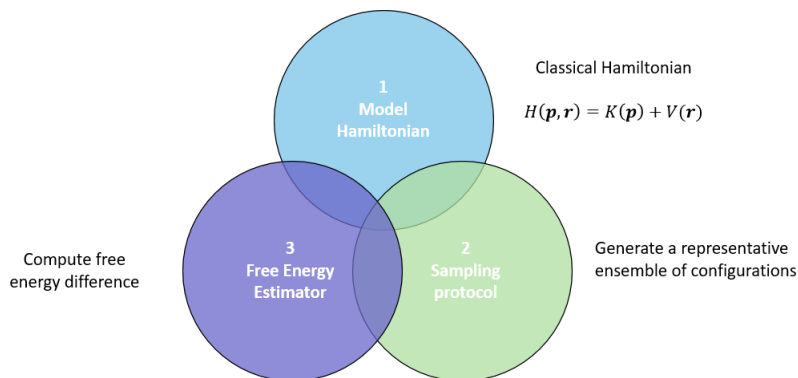


Figure 2.2: Graphical representation of the 3 basic components required for free energy estimation: The model Hamiltonian, The Sampling Protocol, and the Free Energy Estimator

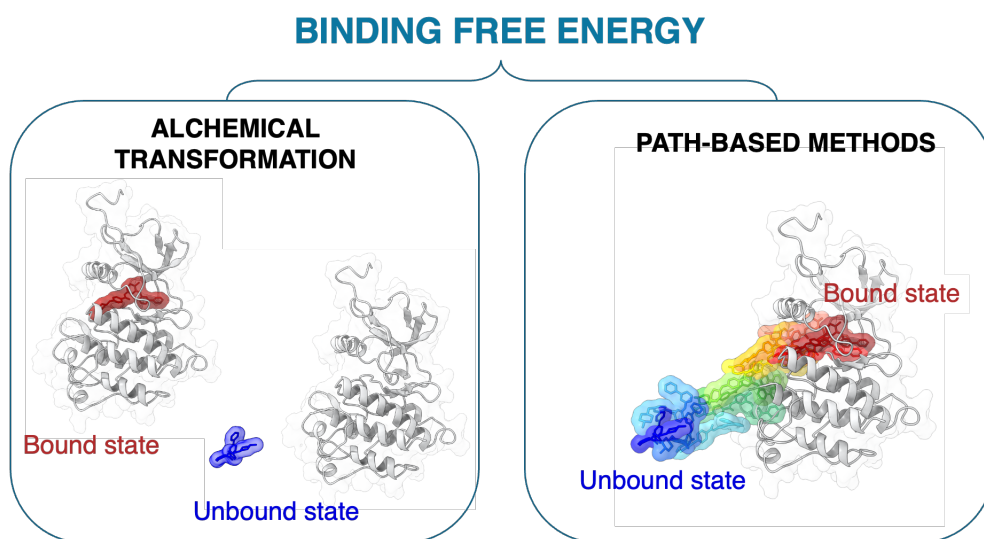


Figure 2.3: Graphical comparison of the two primary approaches for sampling binding events: alchemical methods and path-based methods. In this illustration, the bound state is shown in red, and the unbound state in blue.

2.5.1 Sampling Protocol

As highlighted in Eq. 2.52, the challenge in estimating ΔF_b lies in computing the ratio of the partition functions for the two states of interest. This requires sampling exhaustively all important molecular configurations in both states, which is often difficult with standard MD. Enhanced sampling methods allow to achieve this by improving the sampling and making the estimation feasible.

As mentioned in Section 2.4, to exhaustively sample binding events and computing ΔF_b , two main approaches are commonly used: (i) alchemical methods, which involve nonphysical transformations of the Hamiltonians of the endpoint systems, and (ii) path-based methods, which focus on physical (un)binding pathways. Both approaches are valid, as free energy is a state function and depends only on the initial and final states. Figure 2.3 provides a visual comparison of the two approaches, emphasizing the differences in sampling behavior between them.

Alchemical methods

Alchemical methods are often used to estimate relative binding free energies, while their application for absolute calculations is more complex and less common. These methods involve sampling system transformations through non-physical intermediate, which means they do not provide mechanistic insight into the molecular processes being studied. In 1935, John Kirkwood laid the foundation for two fundamental methods: thermodynamic integration (TI) and free energy perturbation (FEP) [111]. These methods remain nowadays valuable approaches for calculating ΔF_b from alchemical MD simulations.

In general, alchemical transformations involve a hybrid Hamiltonian, typically defined as a linear interpolation between the potential energies of the two specific states (A and B) via the coupling parameter λ . This coupling parameter is $0 \leq \lambda \leq 1$, where $\lambda = 0$ corresponds to the initial state A, while $\lambda = 1$ to the final state B.

$$U(\mathbf{r}; \lambda) = (1 - \lambda)U_A(\mathbf{r}) + \lambda U_B(\mathbf{r}) \quad (2.53)$$

Alchemical transformations involve sampling the system at several different values of λ , employing stratification strategies. This sampling allows for the use of either TI or FEP methods to calculate the free energy difference.

Particularly, in thermodynamic integration, the derivative of the free energy with respect to the continuous parameter λ can be expressed as an equilibrium average over configurations corresponding to a specific λ :

$$\frac{dF}{d\lambda} = \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda} \quad (2.54)$$

The free energy difference between state A ($\lambda = 0$) and state B ($\lambda = 1$) is evaluated by integrating the free energy derivatives along λ :

$$\Delta F_{AB} = \int_{\lambda=0}^{\lambda=1} \frac{dF}{d\lambda} d\lambda = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial U_{\lambda}}{\partial \lambda} \right\rangle_{\lambda} d\lambda \quad (2.55)$$

On the other hand, when using free energy perturbation method, free energy dif-

ferences are computed through the following formula:

$$\Delta F_{AB} = -\frac{1}{\beta} \ln \langle \exp(-\beta \Delta U_{BA}) \rangle_A^{eq} \quad (2.56)$$

where the $\langle \cdot \rangle_A^{eq}$ denotes the equilibrium average of the potential energy difference between the two states, calculated in state A. To accurately estimate free energy differences with these methods, the staging windows at different λ values must be narrow enough to allow the overlapping of distribution between subsequent λ values to facilitate convergence.

In the CADD framework, alchemical transformations are primarily used to estimate the relative binding free energies between two compounds ($\Delta\Delta F_b$), such as ligands A and B within the same congeneric series binding to the same target:

$$\Delta\Delta F_b = \Delta F_b(B) - \Delta F_b(A) \quad (2.57)$$

Here, ligand A is gradually transformed into ligand B in both states: bound to the protein and in the bulk. Then, a thermodynamic cycle can be used to estimate $\Delta\Delta F_b$ [11].

Despite their strong theoretical foundations, alchemical methods have some limitations. Achieving convergence can be difficult, and these approaches do not offer mechanistic or kinetic insights into the process under investigation, as they involve sampling intermediate states that lack direct physical meaning. Consequently, starting from the late 70s, several groups have been devoted to develop innovative algorithms for free energy calculations relying on physical trajectories.

Path-based methods

In contrast to alchemical approaches, physical path-based methods not only estimate free energy differences but also offer detailed insights into the binding pathway itself. These methods sample the system along defined physical paths connecting the bound and unbound states, yielding both the binding free energy and the PMF. Conceptually, the PMF represents the effective potential generated by the average force acting along the selected coordinate or CV [112]. The term PMF arises from

the fact that one way to obtain it is by integrating the mean force acting on the system along a particular path [137]. It captures the underlying free energy variations along the reaction coordinate, and its derivation is:

$$F(\xi) = - \int \langle f(\xi) \rangle d\xi \quad (2.58)$$

where ξ is the model reaction coordinate, and $\langle f(\xi) \rangle$ is the average force (mean force) experienced at the ξ values. As a result, path-based methods can reveal the PMF, along with kinetic details and atomistic-level interactions of protein-ligand binding events, offering valuable insights for rational drug design.

Over time, a variety of path-based enhanced sampling techniques have been developed, with prominent examples including US and MetaD, as discussed in Section 2.4. Additionally, a more in-depth analysis of path-based methods applied to drug discovery, with a specific focus on contributions from our group, will be presented in Section 3.

In the study reported in this thesis, several path-based methods employing geometrical CVs (as PCVs) were utilized to reconstruct the PMF and finally computing the protein-ligand ΔF_b .

2.5.2 Free Energy estimators

The last ingredient to perform free energy calculations is the statistical free energy estimator. Traditionally, specific sampling protocols have been linked to particular estimation methods; however, multiple estimators can often be applied to a single dataset.

Free energy estimators can basically be classified into two different families: (i) equilibrium free energy estimators [138] and (ii) nonequilibrium ones [139]. Moreover, free energy estimators can utilize either unidirectional or bidirectional approaches. Unidirectional methods rely on data collected from simulations that proceed in a single direction, either from the initial to the final state or vice versa (forward/backward simulations). In contrast, bidirectional methods incorporate

data from both starting and target distributions, effectively combining information from simulations that move forward as well as those that reverse from the final state back to the initial one.

Equilibrium Free Energy Estimators

Equilibrium free energy estimators are based on statistical averages obtained from equilibrium MD simulations, where sampling is performed over ensembles distributed according to the Boltzmann distribution. In this context, a classical example is Free Energy Perturbation, which uses the Zwanzig formula [140] to estimate free energy differences between two states from configurations sampled in only one of them. As a one-sided estimator, FEP is widely used in alchemical transformations (Eq. 2.56). However, one-sided methods such as FEP often face convergence issues when there is insufficient phase-space overlap between the two states involved. This lack of overlap leads to poor statistical efficiency and large errors. To overcome these limitations, more advanced two-sided estimators have been developed. These approaches leverage information from both endpoints (or multiple intermediate states) and generally provide more accurate and reliable estimates of free energy differences, particularly in systems where state overlap is limited.

In 1976, Bennett introduced a statistically optimal two-sided estimator that combines information from both forward and reverse simulations to compute free energy differences with minimal variance. This approach, known as the Bennett Acceptance Ratio (BAR) method [141], is commonly used due to its superior efficiency compared to one-sided estimators like FEP. The BAR method can be formulated as:

$$\Delta F_{AB} = -\frac{1}{\beta} \ln \frac{\langle \exp(-\beta \Delta U_{BA}) \rangle_A^{eq}}{\langle \exp(-\beta \Delta U_{AB}) \rangle_B^{eq}} \quad (2.59)$$

where the brackets denote the equilibrium average of the potential energy difference between the two states (A and B), evaluated in the two states separately. The BAR approach is a minimum-variance, asymptotically unbiased estimator for computing free energy differences between two thermodynamic states. In addition to its statistical optimality, BAR can also be derived from a maximum likelihood per-

spective [142]. Under this formulation, BAR determines the free energy difference that maximizes the likelihood of observing the data obtained from both forward and backward simulations, providing an efficient approach to combine information from both ensembles.

Nonequilibrium Free Energy Estimators

Nonequilibrium estimators, in contrast to equilibrium-based methods, derive free energy differences by averaging data collected during nonequilibrium processes. During nonequilibrium simulations, the system is externally driven away from equilibrium along a selected order parameter according to a specified transformation protocol. As the system progresses along this path, the external work required to transition from the initial to the final state is accumulated. The work recorded during these irreversible simulations is called Jarzynski work (W_J). It is computed as the path integral of $\dot{\epsilon} \partial H_\epsilon / \partial \epsilon$ along the trajectory Γ_t [143, 144]:

$$W_J = \int_0^{t_s} dt \dot{\epsilon} \frac{\partial H_\epsilon}{\partial \epsilon}(\Gamma_t) \quad (2.60)$$

Here, ϵ (analogous to λ in thermodynamic integration, but evaluated at equilibrium) regulates the weighting of the additional, time-varying component of the modified potential relative to the original one. W_J represents the cumulative change in the system’s Hamiltonian over time only. Specifically, in the context of binding and unbinding simulations, the computed W_J accounts for the external energy required to drive the ligand from the bound to the unbound state or vice versa (along a predefined pathway). This work reflects the irreversible nature of the transformation and is central to nonequilibrium free energy estimations.

According to the second law of thermodynamics, W_J value will, on average, exceed the corresponding equilibrium free energy difference ΔF by an amount defined as the dissipated Jarzynski work, which quantifies the system’s irreversibility:

$$W_{diss} = \langle W_J \rangle - \Delta F \geq 0 \quad (2.61)$$

This inequality emphasizes that any deviation from quasistatic (reversible) condi-

tions introduces energetic losses, which must be considered when estimating free energy from nonequilibrium trajectories. In practice, during nonequilibrium simulations, the greater the pulling speed, the higher the dissipated work W_{diss} . On one hand, fast simulations are desirable to make efficient use of parallel high-performance computing resources. On the other hand, excessively rapid trajectories inherently lead to greater deviations from equilibrium, thereby increasing dissipation and diminishing the accuracy of free energy estimates. As a result, the optimal balance between simulation speed (i.e., computational efficiency) and accurate convergence of free energy estimates must be defined.

Non equilibrium approaches leverage statistical mechanics theorems, as the Jarzynski equality or the Crooks fluctuation theorem, to extract equilibrium free energies from nonequilibrium trajectories. In 1997, Jarzynski established an exponential relationship between the external work W_J and the equilibrium free energy difference between the states A and B of the system [145]. This estimator is named the Jarzynski equality (JE):

$$\exp(-\beta\Delta F_{AB}) = \langle \exp(-\beta W_J^f) \rangle_{A \rightarrow B} \quad (2.62)$$

where $\beta = 1/k_B T$, and ΔF_{AB} is the free energy difference between the initial and finale states (A and B) of the system under investigation. The angular brackets indicate an exponential average over a number of trajectories N , where the system transitions from its initial state to its target state ($A \rightarrow B$), following the same protocol (i.e., the same variation of parameters over time following the same pathway $A \rightarrow B$). As such, the Jarzynski equality is a one-sided, unidirectional estimator. Since JE involves an exponential average over nonequilibrium work values, it is particularly sensitive to convergence issues. This sensitivity arises because rare trajectories with low work values highly influence the average, a phenomenon often referred to as "rare event dominance." To mitigate this difficulty, a second-order cumulant expansion can be employed as an approximation of the exponential average. This approach assumes that the distribution of work values is approximately Gaussian, an assumption that is generally valid for small perturbations

or near-equilibrium processes. Building on the JE equality, as introduced in Section 2.4.2, Park and Schulten proposed in 2004 a methodology to derive the PMF from nonequilibrium, path-based simulations such as SMD [133]. Their approach establishes a connection between irreversible dynamics and the equilibrium PMF by leveraging the statistical framework provided by JE. In their study, multiple independent replicas of the same steering protocol were executed, and the resulting nonequilibrium work values were used to reconstruct the PMF with high accuracy.

Even if the JE has been extensively applied for free energy calculations, including both alchemical transformations [146] and path-based approaches [147], it has been shown to encounter challenges with statistical convergence when estimating free energy differences. To improve the reliability and accuracy of these estimates, bidirectional nonequilibrium estimators can be employed. A prominent example of such an approach is provided by the Crooks Fluctuation Theorem (CFT) [148]. Analogously to BAR, the CFT combines data from both forward and backward transformations:

$$\frac{P_f(W_J)}{P_b(-W_J)} = \exp(\beta(W_J - \Delta F_{AB})) \quad (2.63)$$

where $P_f(W_J)$ represents the forward W_J distribution, which is the distribution of the Jarzynski work performed as the system evolves from state A to B . On the other hand, $P_b(-W_J)$ refers to the backward W_J distribution, which is the distribution of the Jarzynski work performed as the system evolves from state B to A . Graphically, the free energy difference corresponds to the intersection point of the forward $P_f(W_J)$ and backward $P_b(-W_J)$ distributions. Figure 2.4 illustrates this concept with a focus on binding and unbinding simulations. As depicted, consistent with the second law of thermodynamics, the Jarzynski work values are always greater than or equal to the free energy difference. The averages $\langle W_J^f \rangle$ (highlighted in pink in Figure 2.4) and $\langle -W_J^b \rangle$ (highlighted in light blue) represent bounds for ΔF . By combining the forward and backward transformations, the free energy difference $\Delta \hat{F}_{CFT}$ is constrained by the forward and backward Jarzynski estimates, $(\Delta \hat{F}_f \text{ and } \Delta \hat{F}_b)$, which generally hold as $\Delta \hat{F}_f \leq \Delta \hat{F}_{CFT} \leq \Delta \hat{F}_b$. However, this is not always strictly required [149].

While the BAR estimator focuses on differences in potential energy, the CFT works

with W_J (or more generally, entropy production). Despite these differences, from a mathematical perspective, the two methods are essentially the same algorithm [150]. This parallels the relationship between the Zwanzig and Jarzynski estimators, which are both particular cases of the CFT [142, 151]. As a result, like the BAR method, also the CFT approach can be expressed as a maximum likelihood estimation method [142]. Iterations of the BAR estimator, or its refinements [151], can be applied to solve Eq. 2.63 and compute the free energy difference between two states of the system. When an equal number of forward and backward MD simulations are run, ΔF_{AB} can be estimated by self-consistently solving the reported equation:

$$\langle \{1 + \exp[\beta(W_J^f - \Delta F_{AB})]\}^{-1} \rangle_{A \rightarrow B} = \langle \{1 + \exp[-\beta(W_J^b - \Delta F_{AB})]\}^{-1} \rangle_{B \rightarrow A} \quad (2.64)$$

In recent years, non-equilibrium estimators have been increasingly used in free energy calculations, both for alchemical transformations [146, 152] and path-based approaches [147]. A relevant example of nonequilibrium alchemical transformations is the work done by De Groot’s group. In their method, equilibrium simulations are performed for the two physical end states, while nonequilibrium simulations are employed to bridge them through alchemical space. The JE or CFT is then applied to calculate the free energy differences between these end states [153]. However, to date, the CFT has been predominantly applied in alchemical transformations [146], although its applicability extends to physical collective variables as well [154]. In this context, my research thesis focuses on employing CFT in combination with nonequilibrium path-based methods.

In the end, Figure 2.5 presents a summary of the presented free energy estimators used for binding free energy computation, categorized based on their equilibrium or nonequilibrium nature, as well as their unidirectional or bidirectional approach.

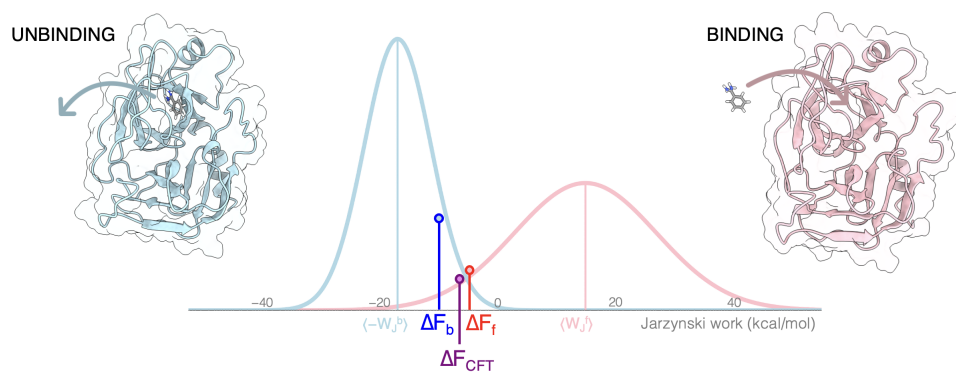


Figure 2.4: Graphical representation of the CFT, where ΔF corresponds to the intersection point of the forward and backward work distributions. The color scheme associates binding events with pink and unbinding events with light blue, providing a clear visualization of how CFT can be applied to estimate binding free energies [P1].

BINDING FREE ENERGY ESTIMATORS

EQUILIBRIUM ΔF ESTIMATORS	NON-EQUILIBRIUM ΔF ESTIMATORS
FEP: Free Energy Perturbation $\Delta F_{AB} = -\frac{1}{\beta} \ln \langle e^{-\beta \Delta U_{BA}} \rangle_A$	1-sided estimators JE: Jarzynski Estimator $\Delta F_{AB} = -\frac{1}{\beta} \ln \langle e^{-\beta W_{A \rightarrow B}} \rangle_{A \rightarrow B}$
BAR: Bennett Acceptance Ratio $\Delta F_{AB} = -\frac{1}{\beta} \ln \frac{\langle e^{-\beta \Delta U_{BA}} \rangle_A}{\langle e^{-\beta \Delta U_{AB}} \rangle_B}$	2-sided estimators CFT: Crooks Fluctuation Theorem $\frac{P_f(W_{A \rightarrow B})}{P_b(-W_{B \rightarrow A})} = e^{\beta(W - \Delta F_{AB})}$

Figure 2.5: Symmetric overview of free energy estimators classified according to their equilibrium or nonequilibrium nature and whether they follow a unidirectional or bidirectional approach.

Chapter 3

On free energy calculations in drug discovery [P5]

Despite the strong theoretical foundation outlined in the previous section and significant methodological advances, binding free energy calculations still face substantial challenges; particularly in drug discovery, where both reliability and computational efficiency are essential. To better frame these challenges and trace the historical evolution of free energy methods, we present a comprehensive overview of our recent account. This work offers both a critical and analytical assessment of the principal methodologies, with special emphasis on their relevance to our laboratory’s research. The focus is placed on their applicability to protein–ligand binding in the drug discovery context, with particular attention to enhanced sampling strategies and both equilibrium and nonequilibrium estimators. The account concludes by outlining the key open questions that continue to define the frontiers of binding free energy research.

3.1 Introduction

In the drug discovery framework, lead compounds (i.e., potential drug candidates) are chemically optimized molecules that exhibit desirable biological activity and are selected for further development along the discovery pipeline. These leads are

often ranked based on several key factors, one of the most important being their binding affinity for the target biomolecule, such as a protein involved in diseases. High affinity generally indicates stronger and more specific interactions with the target, which is crucial for therapeutic effectiveness [155].

With the increasing availability of computational power and the advancement of GPU technologies, *in-silico* methods for predicting and estimating binding affinities have become essential tools in lead optimization [4]. These methods allow researchers to efficiently screen and refine large libraries of compounds before moving to costly and time-consuming experimental testing. The binding affinity of a potential drug to its target is typically expressed in terms of the dissociation constant K_D (Eq. 1.5), which provides a measure of how tightly the compounds binds to the target. This constant is directly linked to the binding free energy through established thermodynamic relationships (Eq. 1.3 and Eq. 1.4).

The presented account provides a comprehensive overview of recent developments and ongoing challenges in the field of protein-ligand binding free energy calculations, with particular emphasis on alchemical and path-based approaches. Conducting this in-depth literature review was a critical component of my PhD work, as it enabled me to thoroughly understand the theoretical foundations underlying these computational methods. In the following, an analytical and critical review of the historical evolution of these methods is reported, highlighting key milestones, methodological advancements, and current limitations. Special attention is given to the contributions and innovations developed within our research group, which have played a significant role in advancing the state of the art. While the theoretical formulations and mathematical background are covered in Section 2.5, this section emphasizes the practical development of the field and contextualizes our group’s work within this scientific landscape.

3.2 Alchemical Transformations in Drug Discovery

As anticipated, the theoretical foundations of free energy calculations were established several decades ago, beginning with the seminal work of Kirkwood in 1935 [111]. Early approaches relied on analytical frameworks, notably the introduction of FEP by Zwanzig in 1954, which was based on perturbation theory [140]. As computational power increased and algorithms based on statistical mechanics matured, it became possible to perform free energy calculations using numerical simulations [105]. A landmark example of this transition is the work by Jorgensen in 1985, who applied FEP in combination with MC simulations to compute hydration free energies [156].

In the context of alchemical transformations (Section 2.5.1), FEP and TI are among the most widely adopted and reliable techniques for computing binding free energies through all-atom MD simulations. Alchemical simulations estimate free energy differences by modeling nonphysical transitions between states, leveraging the fact that free energy is a state function and thus path-independent.

In the drug discovery field, alchemical transformations are commonly employed to calculate the relative binding free energies between analogous compounds, A and B, using the formula $\Delta\Delta F_b = \Delta F_b(B) - \Delta F_b(A)$. In this process, ligand A is gradually transformed into ligand B, both in complex with a protein and in the bulk solvent. By utilizing a thermodynamic cycle, the relative binding free energy $\Delta\Delta F_b$ can be estimated [11].

The first successful calculation of relative binding free energies using this approach was performed by McCammon and colleagues in 1984 [157]. Following this milestone, the application of FEP to ligand-protein complexes became feasible, with major contributions from McCammon, Kollman, and Merz. Today, this method remains widely used in the pharmaceutical industry for ranking therapeutic compounds with similar chemical structures based on their affinity for the target, mostly during lead optimization to prioritise ligands [158, 159].

Subsequently, in 1987, Jorgensen demonstrated the use of FEP to reconstruct free energy profiles along a defined coupling parameter [160]. Building on this foundation, more advanced thermodynamic cycles were later developed to estimate the absolute binding free energy. These alchemical approaches involve transforming the ligand into a non-interacting particle, effectively decoupling it from both the protein and the solvent environment. First introduced by Jorgensen [160], this method was later rigorously developed by Gilson in 1997 [161], and is now widely known as the double decoupling method. Figure 3.1 reports a schematic representation of a thermodynamic cycle based on alchemical double decoupling method.

Despite their strong theoretical basis, alchemical methods face practical limitations, particularly in terms of convergence. FEP, being an equilibrium and unidirectional estimator, relies on sampling equilibrium configurations from a reference state (A) to a target state (B). However, for complex molecular systems, the potential energy distributions of the initial and final states often show limited overlap, leading to poor convergence [139]. To address this, simulations are typically stratified into multiple intermediate states with coupling parameters $0 < \lambda < 1$, improving overlap between neighboring states and enhancing convergence. Nevertheless, even with stratification, sufficient overlap between adjacent λ windows is essential for reliable results. Importantly, convergence in the forward direction ($A \rightarrow B$) does not guarantee convergence in the reverse ($B \leftarrow A$), reflecting the inherent directionality of FEP.

To improve the accuracy and efficiency of free energy difference estimations, Bennett introduced in 1976 a statistically optimal method known as BAR [141]. This approach combines data from both forward and backward transformations, significantly reducing variance compared to unidirectional estimators like FEP. Bennett also demonstrated that performing an equal number of forward and backward simulations yields results close to the statistical optimum. Nearly three decades later, in 2003, the Bennett Acceptance Ratio method was rigorously rederived as a maximum-likelihood estimator for computing free energy differences, offering improved statistical efficiency [142]. As a result, many modern alchemical free energy protocols incorporate bidirectional sampling strategies using BAR to enhance the

reliability and precision of their results [162, 163]. Initially, alchemical transformations were predominantly applied to small, structurally similar ligands, as large chemical perturbations tend to hinder the convergence of FEP calculations. Moreover, special care is required when dealing with charged ligands, particularly in accurately accounting for electrostatic decoupling, which can significantly affect the reliability of the results [162]. Despite these challenges, alchemical methods have evolved considerably and are now capable of handling more complex systems. Notably, they have been successfully applied to compute binding free energies for large, flexible ligands of pharmaceutical relevance. A prominent example is the cancer drug Gleevec, whose binding to two kinase targets (Abl tyrosine kinase and c-Src) was explored using alchemical approaches [164].

In addition to FEP, TI has also been successfully employed in drug discovery and to estimate solvation free energies, a key property in drug development due to its strong influence on oral bioavailability and absorption [165]. In our research group, TI was successfully employed to assess the solubility advantage of amorphous ketoprofen over its crystalline counterpart [166]. This study introduced a novel approach combining harmonic dynamics with MD simulations to calculate the absolute free energy of transferring a ketoprofen molecule from the condensed phase to solution. This was achieved by summing a reference free energy (including analytically derived ideal and harmonic terms) with anharmonic corrections obtained through TI. Two thermodynamic legs were used: one connecting the harmonic and anharmonic Hamiltonians at low temperature, and another extending the results to 300 K. The calculated free energies closely matched experimental data, confirming the enhanced solubility of the amorphous form.

Since their successful application in the lead optimization of anti-HIV-1 agents by Jorgensen in 2005 [159], alchemical methods have remained a widely used and trusted approach for calculating binding free energies in the pharmaceutical industry. Moreover, alchemical methods are not limited to equilibrium scenarios. A notable example of nonequilibrium alchemical transformations is the work by De Groot and colleagues, which demonstrated the efficiency of such approaches in binding free energy calculations [153].

However, the core strength of alchemical methods also represents a key limitation. Because these approaches rely on nonphysical intermediates to interpolate between molecular states, they do not reflect real dynamical pathways. As a result, alchemical transformations offer no direct information about the mechanism or kinetics of the binding process, limiting their ability to provide insight into how or how quickly a ligand interacts with its target.

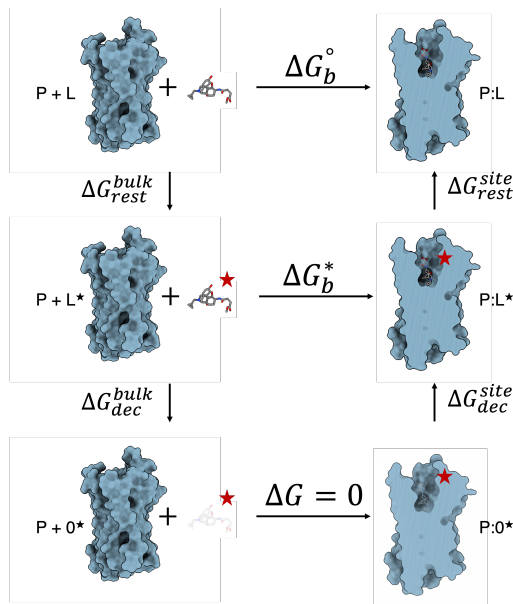


Figure 3.1: Schematic representation of a thermodynamic cycle based on alchemical double decoupling for binding free energy calculation. To prevent the ligand (L) from drifting away from the protein (P) binding site during decoupling, positional, orientational, and conformational restraints are applied (). The restrained ligand is then decoupled from both the protein and the bulk solvent (denoted by 0). The free energy contribution associated with introducing these restraints is explicitly evaluated, and, by following the thermodynamic cycle, the overall binding free energy can be determined.

3.3 Path Based methods in Drug Discovery

In the drug discovery context, free energy calculations relying on physical trajectories are particularly challenging, as many interesting biochemical processes (including protein-ligand binding) are rare events. Thus, the time and costs required to sample these events are beyond the computational resources required to routinely apply MD simulations for rational drug design. To effectively apply simulations in

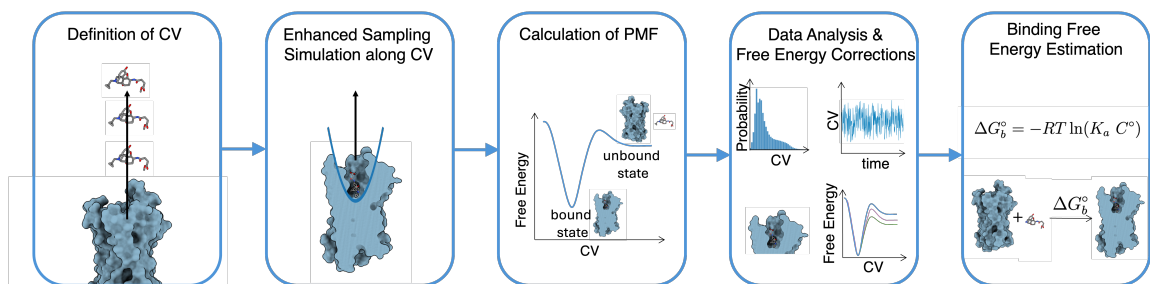


Figure 3.2: Schematic representation of a workflow for calculating binding free energy using path-based methods. First, an optimal CV is identified. Then, enhanced sampling methods are used along the chosen CV. From these simulations, and with free energy estimators, the PMF along the CV is calculated. Successive analysis include convergence analysis, visual inspection of trajectories, and definition of pipeline-specific free energy corrections. In the end, the binding free energy is computed.

drug discovery, enhanced sampling techniques are essential. Among the most established path-based approaches there are Umbrella Sampling, Metadynamics, and Steered Molecular Dynamics. A schematic representation of a generic path-based workflow is shown in Figure 3.2.

3.3.1 Umbrella sampling

Since its introduction in 1977 [118], US has become a well-established method for investigating ligand–receptor dissociation and computing binding free energies. The core concept of US is to apply a biasing potential to enhance sampling along specific CVs, particularly in regions of phase space that are otherwise rarely visited. This approach allows for more efficient exploration of transition states and high-energy regions. To recover unbiased Boltzmann-weighted distributions and the corresponding PMF, several reconstruction techniques have been developed, including the Weighted Histogram Analysis Method [119, 120], Umbrella Integration [167], and the Dynamic Histogram Analysis Method [168].

In an early work by Patey and Valleau, US with stratification was used to reconstruct the free energy profile of an ion pair in a polar fluid [169]. Subsequently, in the '80s, Karplus applied US to calculate free energy differences in biological systems [169] and later improved the method by developing an adaptive US approach, with the WHAM serving as its core equation [170]. Additionally, Jorgensen was

among the early adopters of US, using it to study the conformational equilibria of n-Butane in dilute solution [171].

In recent years, US has found broad application in areas such as ligand-binding affinity estimation and the investigation of complex reaction mechanisms.

A significant advancement was presented in 2011 by Branduardi *et al.* [172], who proposed a novel strategy for exploring reaction mechanisms. They integrated US with SMD using PCVs to construct a good-resolution free energy landscape of the reaction pathway. This approach, implemented within a hybrid QM/MM DFT framework, provided a detailed and physically meaningful description of the process. In 2012, Lodola and coworkers [173] employed free energy simulations based on US within a QM/MM framework to investigate the reaction mechanism of amide hydrolysis. They estimated the free energy involved in the catalytic process using an iteratively optimized path guided by SMD with PCVs. Some years later, Di Martino *et al.* (2014) [174] investigated the cis-trans isomerization of a model peptide system using US simulations in both the Pin1-bound and unbound states (peptidyl-prolyl isomerase Pin1). Their study provided free energy barriers that were consistent with experimental data.

Moreover, in our laboratory, US has also been applied to compute binding free energies. In the work by Decherchi *et al.* (2015) [175], US simulations were performed using PCVs to investigate the binding of the transition state analogue DADMe-immucillin-H to the purine nucleoside phosphorylase enzyme. To define the initial binding pathways, standard MD simulations were combined with a machine learning-based analysis. These pathways were analyzed using US, and the resulting biased data were processed through a regularized nonlinear regression technique to reconstruct the free energy landscape.

The nonlinear regression technique employed in that work was a refined version of the method introduced by Maragliano and colleagues [176]. In [175], the mean force along the reaction coordinate was smoothly interpolated using a Regularized Least Squares algorithm combined with a Gaussian kernel, allowing for flexible, non-parametric fitting. To ensure numerical stability and avoid overfitting, a Tikhonov regularization term was included. Also the width of the Gaussian was kept rela-

tively large to further reject noise. This strategy resulted in a smooth free energy profile.

However, US-based strategies may be very elaborate and require the identification of two critical parameters: the windows in the CV space to perform the simulations and the force constants for the harmonic restraints. Due to these complexities, methods based on US are not routinely used in drug discovery. Overall, US is the first of a series of path-based enhanced sampling methods that aim to accelerate rare events. Several other enhanced sampling strategies have been proposed, such as those involving higher temperatures or external potentials

3.3.2 Metadynamics

One of the most widely used enhanced sampling techniques for accelerating rare events is MetaD, including its refined variant WT-MetaD [124, 125], along with several related metadynamics-based strategies [114]. Since its introduction, WT-MetaD has been applied across a broad spectrum of scientific disciplines due to its ability to efficiently explore complex free energy landscapes. A pioneering application in the context of ligand binding was carried out by Gervasio *et al.* [128], who demonstrated its potential for investigating protein-ligand interactions. Today, MetaD and its variants are routinely employed in drug discovery to elucidate binding mechanisms in detail and to compute PMF profiles, offering mechanistic and thermodynamic insight into molecular recognition processes.

In recent years, our research groups have widely employed metadynamics and metadynamics-related methods [128, 177, 178, 179] in the context of rational drug design. Particularly, in 2009 well-tempered MetaD [178] was employed to refine docking poses, accounting for molecular flexibility and exploring binding mechanisms. According to the authors, refining docking poses with well-tempered MetaD can be as straightforward as refinement via umbrella sampling approach. This successful protocol not only enhances the predictive accuracy of docking but also reveals the full putative unbinding mechanisms of compounds of interest. A detailed overview of metadynamics-based algorithms and their application in drug discovery, particularly for exploring ligand-protein binding and unbinding mechanisms, is

provided in a review from our laboratory [114]. While these methods have proven highly effective in mechanistic studies, their application has largely been confined to retrospective analyses. In contrast, alchemical free energy methods remain the standard approach for prospective drug design due to their robust performance in predicting binding affinities. Nonetheless, MetaD holds significant potential within drug discovery pipelines. Beyond thermodynamic estimations, MetaD can reveal detailed mechanistic insights into binding and dissociation processes [11], which are critical for the rational optimization of lead compounds and the identification of kinetic bottlenecks or metastable states along binding pathways.

Attempting to improve the accuracy and automation of MetaD-based approaches for computing binding free energy differences, Bertazzo *et al.* [180] developed an advanced protocol that integrates enhanced sampling techniques, machine learning, and specialized optimization algorithms. This framework was designed to reduce both computational cost and the number of free parameters typically required in free energy calculations, thereby streamlining the simulation setup and analysis process. The protocol was successfully applied to glycogen synthase kinase 3 beta (GSK-3 β), a protein kinase of pharmacological relevance, and achieved excellent agreement with experimental binding affinities, both in relative and absolute terms. This work represents the state of the art in MetaD methodologies within our laboratory.

3.3.3 Steered Molecular Dynamics

Another enhanced sampling technique based on the application of biasing potentials is SMD, which was developed in the 1990s [129, 130]. SMD was designed to probe rare biological events that are otherwise difficult to access through conventional MD. Notable early applications include studies on protein folding [131] and protein-ligand binding [132].

The first application of SMD in the context of drug discovery from our group was presented by Colizzi *et al.* in 2010 [134]. In this pioneering study focused on antimalarial drug development, SMD was employed to investigate the binding properties of flavonoid compounds to their target enzyme. By analyzing the

force profiles generated during ligand unbinding simulations, the method was able to distinguish active inhibitors from inactive analogues. To quantitatively assess the binding strength, potential of mean force profiles were reconstructed using the second-order cumulant expansion of the Jarzynski equality [145]. This allowed for a detailed computational evaluation of ligand affinity without requiring equilibrium sampling. The unbinding pathways explored in the study were initially identified using random acceleration molecular dynamics [181], providing plausible dissociation routes for SMD analysis. Overall, this work demonstrated the potential of SMD as a structure-based drug design tool capable of revealing key molecular interactions and supporting ligand optimization efforts in a realistic, fully solvated environment.

Further contributions from our group to the application of SMD in drug discovery are exemplified by the studies of Patel *et al.* [182, 183]. Building on the foundation laid by Colizzi *et al.*, Patel and colleagues introduced a novel analysis framework for interpreting nonequilibrium SMD trajectories [182]. Rather than relying on the JE [145], which can be computationally demanding for accurate PMF reconstruction, their approach focused on extracting meaningful structural and mechanistic information directly from a small set of pulling simulations. This post-processing strategy demonstrated that even without achieving full convergence or applying the JE, SMD work profiles contain rich data that reflect key conformational transitions and interactions critical to the binding or unbinding process. The method highlights how nonequilibrium simulations can still yield practical insights into ligand-receptor dynamics while significantly reducing computational overhead.

Building upon the optimized SMD setup developed in the earlier work, Patel and colleagues later extended the methodology to explore the unbinding mechanisms of nine kinase inhibitors [183]. In this study, ligands were characterized by analyzing the decay patterns of rupture forces, unbinding times, and reconstructed PMF profiles. The approach proved effective at distinguishing active from inactive compounds within chemically related series, highlighting differences in their interaction strength and binding dynamics. However, the method showed limitations when attempting to rank inhibitors with comparable potencies but divergent chemical

scaffolds. Despite this, the simulations consistently offered valuable atomistic-level insights into the unbinding processes, revealing key interactions and conformational changes. This is particularly relevant given the inherent complexity of kinase targets, which often feature highly flexible and solvent-accessible binding sites that challenge traditional structure-based drug design strategies.

Additional early advances in the application of SMD to computational drug discovery were made by Mai *et al.* in 2010 and 2011 [184, 185]. These studies demonstrated the use of SMD to evaluate and validate the binding capabilities of various ligands, highlighting its practical utility within structure-based drug discovery workflows.

All these contributions marked important progress toward the routine use of SMD in rational drug design. One of the notable advantages of SMD highlighted by these works is the ability to precisely control simulation time through predefined input parameters, offering a degree of predictability and efficiency that is often lacking in alternative enhanced sampling techniques. For instance, while MetaD can yield detailed mechanistic insights, its convergence time is difficult to anticipate and can vary widely depending on the system and the collective variables used. This makes SMD particularly attractive in early-stage screening or hypothesis-driven studies where computational cost and timeline are critical factors.

In the end, the latest advancement in SMD applications within our group is presented in our study [P1], which builds upon the framework introduced in [180]. This work integrates SMD with PCVs to reconstruct the PMF and compute absolute binding free energy differences. The central aim was to explore the potential of combining nonequilibrium simulations with PCVs to enhance the accuracy of binding free energy calculations. Recognizing the convergence limitations associated with the JE [145, 144], the study employed the CFT [186] as an alternative bidirectional estimator. By incorporating both forward and backward trajectories, CFT provides a statistically more robust framework, analogous to the BAR method, which is widely used in equilibrium contexts. The newly developed semi-automated pipeline offers both accurate estimates of absolute binding free energy differences and valuable insights into ligand dissociation mechanisms at the atomic

scale.

Moreover, further investigations from our second work [P2] have underscored the role of dissipation in nonequilibrium free energy calculations, leading to the formulation of practical guidelines aimed at minimizing irreversible work. These methodological advancements have allowed the successful application of the protocol to complex and pharmaceutically relevant systems, including the Abl–Gleevec kinase complex and RNA-ligand interactions, underscoring its versatility and robustness in drug discovery contexts [P3].

3.4 Challenges and perspective

This chapter provides an overview of MD-based strategies and milestone approaches used in drug discovery for estimating binding free energies, with particular focus on research conducted in our laboratory.

In reference [162], authors present two formal approaches for performing binding free energy calculations: one based on alchemical transformations and the other on physical CVs. Both methods utilize similar restraints to enhance sampling convergence, which must be rigorously accounted for, and they are derived within a comparable formal framework. Even if these two approaches are conceptually distinct, they ultimately produce equivalent results [162].

The selection between these methods typically depends on the nature of the system under investigation. Alchemical strategies are particularly effective in high-throughput scenarios, such as the hit-to-lead phase of drug discovery, where ranking a large set of ligands is essential. They are also advantageous when the system lacks well-defined reaction coordinates, as they do not require explicit knowledge of relevant CVs. Nevertheless, alchemical transformations are often limited to small and not highly charged ligands, due to convergence issues resulting from large perturbations and solvation free energies. Moreover, a significant drawback is their inability to provide mechanistic or kinetic insights into the binding process.

Conversely, path-based methods offer detailed information about the transforma-

tion mechanism and kinetics. However, their performance heavily depends on the choice of CVs; inappropriate selections can lead to inefficient sampling and hinder convergence. Despite this limitation, approaches that rely on physical pathways are often the most appropriate for investigating the binding of large ligands, especially in the context of protein–protein interactions.

A compelling alternative to traditional alchemical and physical path-based free energy approaches is the Alchemical Transfer Method (ATM), recently introduced by Gallicchio and colleagues [187]. This method operates by perturbing the ligand position between the receptor’s binding site and the surrounding explicit solvent. Conceptually, ATM bridges the gap between purely alchemical strategies (like double decoupling) and path-based techniques, as it incorporates both non-physical transformations and spatially resolved ligand motions characteristic of physical methods. Moreover, another valuable and innovative strategy is Hamiltonian Replica Exchange MD (H-REMD). In this method, several system replicas are run in parallel, each governed by a differently modified Hamiltonian (often achieved by scaling nonbonded interaction parameters). Periodic exchanges between these replicas are performed to enhance the sampling efficiency of intermediate states, thereby improving the convergence of both alchemical and path-based free energy methods [188, 189, 190]

Overall, as highlighted throughout this account, MD-based methods and free energy calculations offer significant potential as powerful tools for rational drug design. Nonetheless, despite notable progress in recent years, several challenges remain. Key limitations include: (*i*) the inherent difficulty in thoroughly sampling rare events such as protein–ligand binding, (*ii*) limitations in the accuracy and transferability of current FFs, and (*iii*) convergence problems that can affect the reliability and reproducibility of computed free energy values.

Considering the sampling limitations, they arise when high-probability regions in the configuration space are separated by significant energy barriers, as is often the case in protein-ligand binding. Traditional MD methods struggle to effectively sample these regions, making it difficult to capture binding events. This issue, known as quasi-non-ergodicity, highlights the need for enhanced sampling

techniques that are specifically designed to overcome such barriers. A variety of enhanced sampling methods are currently available, each offering unique advantages and limitations. These techniques are designed to facilitate the exploration of the configuration space by overcoming energy barriers, thus enabling more efficient sampling of rare yet critical events. Advances in the development of more sophisticated CVs, including those driven by machine learning, are helping address these sampling challenges. These novel CVs, tailored to drug discovery applications, are improving the efficiency and accuracy of sampling, further enhancing the potential of MD approaches in rational drug design.

The second critical factor to consider is that computer simulations generate microscopic structural ensembles, where the statistical weights are determined by the theoretical model used. In protein-ligand simulations, the theoretical models commonly employed are FFs. As a result, the accuracy of simulation outcomes is directly dependent on the precision of the FF utilized. To overcome these limitations, machine learning-based FFs have been developed, providing a promising approach to enhance the representation of potential energy surfaces. In particular, our group has recently introduced a novel Neural Network Potential (OBIWAN), which holds significant potential for advancing the field by offering more accurate and flexible models for simulating molecular systems [191].

In the end, convergence issues, though often overlooked, can significantly affect the reliability of computational results. These problems can arise in different contexts, such as sampling convergence, which is commonly encountered in MetaD simulations, or computational convergence, which occurs when using specific free energy estimators. Achieving convergence is essential for obtaining robust, reproducible results. For free energy estimators, performing rigorous statistical analyses is crucial to evaluate and validate convergence. For instance, in our work [P1], we presented a detailed study on the convergence of free energy calculations, emphasizing the importance of systematic convergence assessments to ensure the reliability and accuracy of computational results.

In conclusion, all of these recent advancements are paving the way for the broader adoption of MD-based methodologies in drug discovery, enriching the toolkit for es-

timating binding and unbinding free energies. These computational strategies hold the potential to accelerate the identification and optimization of novel bioactive compounds, thereby significantly advancing the drug discovery process.

Building upon the theoretical framework presented thus far, the following chapters detail the practical application of these principles. I present the development and progressive refinement of our novel computational pipeline based on nonequilibrium methodologies. All data and scripts necessary to reproduce our results are publicly available at <https://gitlab.iit.it/hpc/FreeEnergyPath/>. Some scripts and simulations require the BiKi Software, and a temporary license can be requested for free to reproduce our findings. In addition, the concluding chapters explore complementary investigations that continue to hold significance within the broader context of drug discovery.

Chapter 4

Bidirectional Path-based nonequilibrium simulations for binding free energy [P1]

4.1 Aim and Objectives

Driven by the growing interest in nonequilibrium techniques [192, 193, 194, 195] and the advancements in parallel and GPU-based computing architectures, we want to propose a new computational pipeline for calculating binding free energy in the nonequilibrium scenario. This novel approach builds on a previous workflow developed in our laboratory [180] and introduces novel features that facilitate easy parallelization of simulations. As a result, this method based on nonequilibrium simulations and nonequilibrium estimators has the potential to significantly reduce the time needed to obtain accurate estimates of binding free energy, while also providing valuable insights into the molecular-level binding mechanism.

4.2 Introduction

The binding free energy (ΔF_b) of a potential drug to its macromolecular target is a critical metric directly linked to the drug’s affinity. As such, it is a key

parameter in optimizing new drug candidates during the drug discovery process [4, 11]. Estimating binding free energy, however, remains a significant computational challenge, especially in complex biological systems with many degrees of freedom. Indeed, modern computational methods frequently struggle with practical applications in pharmaceutical systems, primarily due to challenges in achieving reliable convergence in free energy estimations.

To address these limitations, enhanced sampling methods can be used. As outlined in Section 2.5.1, these methods generally follow two main approaches: (i) alchemical methods and (ii) path-based methods. The latter not only estimate free energy, but also provide valuable insights into the binding mechanism at the molecular level. Additionally, as explained in Section 2.5.2, statistical free energy estimators used to compute ΔF_b from simulations can be categorized into two categories: (i) equilibrium free energy estimators [138], and (ii) nonequilibrium free energy estimators [139]. While the former approaches are well established and widely used, the latter family has emerged more recently and has gained increasing attention in recent years due to its inherent advantages.

The presented study explores the possibility to apply nonequilibrium free energy estimators within path-based methods for binding free energy estimation. The proposed protocol incorporates novel features that facilitate straightforward parallelization of simulations, potentially leading to a substantial decrease in the computational time needed to achieve accurate binding free energy estimates.

4.3 Methods

Nowadays, various path-based enhanced sampling methods are available to accurately sample rare events and reconstruct the free energy profile related to protein-ligand binding [105]. However, selecting an ideal CV to describe the physical path remains a challenge. Starting from previous research from our laboratory on free energy calculations of protein-ligand interactions [180], we expanded this methodology to a nonequilibrium framework. In this work, we propose replacing the inherently sequential MetaD algorithm with a straightforwardly parallelizable approach

that also offers improved control over convergence. Additionally, we introduce a semi-automated protocol for PMF reconstruction utilizing SMD and PCVs. This protocol allows to estimate the standard protein-ligand binding free energy, which can be compared successively with experimental values.

As highlighted in Section 2.4.2, SMD is a nonequilibrium technique, thus nonequilibrium estimators as the CFT [148, 186] or the JE [144, 145] are required to reconstruct the PMF and calculate the binding free energy. In our pipeline, SMD is applied in combination with PCVs, leading to the following equation for the time-dependent harmonic potential $R(x, t)$:

$$R(x, t) \equiv \frac{1}{2}k(S(x) - \hat{S}(t))^2 \quad (4.1)$$

where $\hat{S}(t)$ represents the pulling coordinate, which advances at a constant velocity throughout the simulations, transitioning from the initial to the final state.

The details of our computational pipeline are reported below, where we provide a thorough explanation of the definition and formulation of the PCVs used, the reconstruction of the PMF from work values, and the calculation of the standard free energy of binding.

Figure 4.2 illustrates the five key steps of this innovative workflow: (*i*) generation of an initial guess path, (*ii*) preparation and optimization of the minimum free energy path, (*iii*) SMD simulations via PCVs, (*iv*) application of nonequilibrium estimators and PMF reconstruction, and (*v*) computation of the standard binding free energy.

4.3.1 Path collective variables and Path definition

The first stage of the pipeline is an exploratory step aimed at identifying an optimized minimum free energy path, which will subsequently be used in combination with PCVs. To generate initial candidate trajectories that explore the unbinding event under investigation, ABMD simulations [135] are employed, using an electrostatic-like collective variable [196, 197, 198]. In this approach, a target value for the CV is set, and a moving harmonic restraint ensures it is reached

within a finite time [135]. A key advantage of this method is its gentle application of restraints, allowing the system to evolve through thermal fluctuations rather than a predefined time schedule, as seen in SMD. To run these simulations, we employ an electrostatics-like CV previously validated in our laboratory [198]. Essentially, this electrostatics-like CV is a Yukawa (or screened-Coulombic) potential that, instead of relying on the original system charges, assigns equal formal charges (opposite in signs) to the ligand and the binding site, respectively. Thus, the resulting electrostatics-based ABMD uses a fictitious electrostatic potential to promote ligand dissociation. Through this approach we guide the systems towards a zero electrostatic-interaction energy between the protein and the ligand, which corresponds to the complete dissociation of the two partners. This method ensures a smooth dissociation of the ligand over time.

From this initial set of unbinding trajectories, the one with the shortest unbinding time, highest frequency of occurrence, and a mechanistically relevant pathway is selected as the starting point for subsequent path optimization. The refinement process of the path involves two sequential algorithms: the principal path algorithm [199] and the equidistant waypoint algorithm [180]. The principal path algorithm, a machine learning-based approach, constructs a smooth transition pathway between the bound and unbound states by selecting a series of representative conformations from MD data. The equidistant waypoint algorithm then refines this path using consecutive short SMD steps, ensuring equally spaced configurations in the path (based on RMSD), which is crucial for the definition of PCVs.

The equations defining the two PCVs, $S(x)$ and $Z(x)$, are provided in Section 2.4.1, specifically in Equations 2.41 and 2.42. As anticipated, PCVs describe the position of the system in configurational space relative to a predefined transition path, providing an effective framework for analyzing ligand unbinding mechanisms. Moreover, a graphical representation of $S(x)$ and $Z(x)$ using a simplified host-guest complex is shown in Figure 4.1. This illustration provides a visual understanding of how these collective variables describe the system’s progression along the binding and unbinding pathway.

4.3.2 Nonequilibrium Simulations and Free Energy Estimation

During 2-sided SMD simulations, PCVs guide simulations along the predefined pathway reducing ambiguities in the process evolution and effectively acting as an implicit restraint. This approach enables the implementation of bidirectional algorithms, where both forward and reverse calculations contribute to the analysis. The SMD harmonic restraint at constant velocity (Eq. 4.1) is applied along the pulling coordinate $\hat{S}(t)$, which tracks movement along the reference path. Additionally, harmonic restraints are imposed to confine the system along $Z(x)$, which are activated only when a specified threshold is exceeded.

During all these irreversible simulations, the Jarzynski work (Eq. 2.60) is computed and then used to reconstruct the PMF. Particularly, for the systems investigated in this study, we conduct 50 independent binding and unbinding simulations using the same protocol, and we collect W_J values to construct work profiles. Moreover, simulations are performed at different pulling speeds to assess how the pulling rate and the number of replicas influence the convergence of free energy estimates.

In nonequilibrium scenario, the observed W_J represents the energy required to transition the ligand between bound and unbound states. On average, this value exceeds the binding free energy by an amount known as the dissipated Jarzynski work (Eq. 2.61). Higher pulling speeds in SMD lead to increased dissipation of work. Thus, balancing computational efficiency and accuracy presents a challenge: while shorter simulations maximize the benefits of parallel high-performance computing, excessively fast trajectories can lead to significant dissipation. In this study, we explore the optimal trade-off between execution speed and the reliable convergence of free energy estimates. Moreover, we analyze the contribution of W_J to the overall calculations while distinguishing its impact from the statistical efficiency and potential bias of the estimators.

Subsequently, nonequilibrium estimators are applied to the W_J values obtained from the simulations to compute the PMF. The protocol relies on both JE and its bidirectional version the CFT, with a maximum likelihood formulation of the

CFT [142]. Practically, with the same number of forward and backward replicas, we solve iteratively the following equation in a self-consistent way:

$$\langle \{1 + \exp[\beta(W_J^f - \Delta F_{AB})]\}^{-1} \rangle_{A \rightarrow B} = \langle \{1 + \exp[-\beta(W_J^b - \Delta F_{AB})]\}^{-1} \rangle_{B \rightarrow A} \quad (4.2)$$

where ΔF_{AB} is the free energy difference between the states A and B of the system ($\Delta F_{AB} = F(B) - F(A)$); and the subscripts f and b denote the forward and backward transformations, respectively.

Considering the PCVs definition in our workflow, we effectively solve the following equation:

$$\langle \{1 + \exp[\beta(W_J^f - \Delta F_{S_i S_{i+1}})]\}^{-1} \rangle_{A \rightarrow B} = \langle \{1 + \exp[-\beta(W_J^b - \Delta F_{S_i S_{i+1}})]\}^{-1} \rangle_{B \rightarrow A} \quad (4.3)$$

where W_J^f and W_J^b are the work values from forward and backward simulations related to the S_i, S_{i+1} interval. The PMF $F(S_{i+1})$ is $F(S_i) + \Delta F_{S_i S_{i+1}}$ and the i index represents the i -th configuration along the path. The PMF is reconstructed by progressively solving Eq.4.3 for increasing values of i . A similar method is applied for the JE too.

4.3.3 Overall Computational pipeline

In summary, our novel nonequilibrium workflow consists of five key steps:

1. Generation of an initial trajectory describing the unbinding event under investigation with ABMD coupled with an electrostatic-like CV [198]. Multiple exploratory simulations are run, and from these preliminary trajectories, the one with the highest frequency and mechanistic relevance is selected as the initial guess path.
2. Identification of a putative minimum free energy path through two path algorithms. First, the unbinding guess path is optimized in the configurational space using the machine learning-based principal path algorithm [199]. This is followed by refinement using an enhanced version of the Equidistant Waypoints Algorithm [180]. The result is a smooth path composed of equidistant

molecular configurations that accurately represent the unbinding process.

3. Running nonequilibrium bidirectional SMD simulations following the reference path via PCVs. During these irreversible simulations, the W_J values are computed and collected to reconstruct the work profiles. Particularly, for each integer value S_i , the W_J corresponding to each replica is computed. Several binding and unbinding SMD simulations are run (following the same transformation protocol) until convergence of the binding free energy estimate is reached.
4. PMF reconstruction through the use of nonequilibrium estimators via the devised automatic procedure. Specifically, JE and CFT are applied to all replica W_J values associated with a given S_i to determine individual points along the PMF profile. This process is repeated for all integer values of S_i , ultimately generating the free energy profile along S .
5. Finally, the standard binding free energy is determined by summing the binding free energy obtained from simulations with the standard volume correction term, as suggested by Doudou *et al.* [200]:

$$\Delta F_b^\circ = \Delta F_b - \Delta F_v = -RT \ln \frac{Q_{site}}{Q_{bulk}} - RT \ln \frac{V_{bulk}}{V^\circ} \quad (4.4)$$

Here, ΔF_b is the ratio of the canonical partition functions of the bound (Q_{site}) and unbound (Q_{bulk}) states, calculated by integrating the PMF:

$$\frac{Q_{site}}{Q_{bulk}} = \frac{\int_{site} \exp\left(-\frac{F(S)}{RT}\right) dS}{\int_{bulk} \exp\left(-\frac{F(S)}{RT}\right) dS} \quad (4.5)$$

To compute this ratio, a reference frame that distinguishes between the two regions is necessary. The discriminating value of $S(x)$ corresponds to a specific molecular configuration along the reference path. This value is determined through an analysis of the PMF, supplemented by visual inspection of the trajectories.

Then, the standard volume correction term (ΔF_v) is computed to obtain

standard estimates that can be compared with experimental values. ΔF_v accounts for the difference in free energy when considering the standard-state volume V° of $1661 \text{ \AA}^3$ (corresponding to a 1 M concentration) instead of the unbound volume V_{bulk} sampled during the simulations. Specifically, we first identify ligand configurations along the reference path that represent the unbound state. These configurations are aggregated, and the solvent-excluded surface is computed over the merged structure. The resulting volume serves as an estimate of the effective unbound space sampled by the ligand. This correction is calculated using NanoShaper [201] and the BiKi software [202].

Notably, all these steps are designed under the assumption of a stiff spring approximation, ensuring a clear correspondence between the instantaneous value of $S(x)$ during the simulation and the time-dependent center of the harmonic potential $\hat{S}(t)$, thereby avoiding any ambiguity.

In the end, the errors associated to our estimates are evaluated using the bootstrapping method [203], which is a standard approach for statistical error estimation. This procedure involves generating multiple virtual datasets by sampling with replacement from the original n samples. The statistical quantity of interest is calculated for each virtual dataset, and the final estimate is obtained by averaging the results across all bootstrap samples. In this study, n trajectories associated with Jarzynski work values are used. Bootstrapping is applied to the work values corresponding to each integer value of the pulling coordinate. The nonequilibrium estimator is then used to compute multiple estimates of ΔF_b , and their average provides both the average binding free energy and its associated uncertainty.

More detailed information on each stage of the pipeline, along with the specifics of system preparation and simulation setups, can be found in reference [P1].



Figure 4.1: Pictorial representation of PCVs for the host-guest system. $S(X)$ quantifies the position along the predefined path connecting the bound and unbound states. $Z(X)$ measures the deviation of actual trajectories from this reference path, capturing how far the system strays from the intended route.

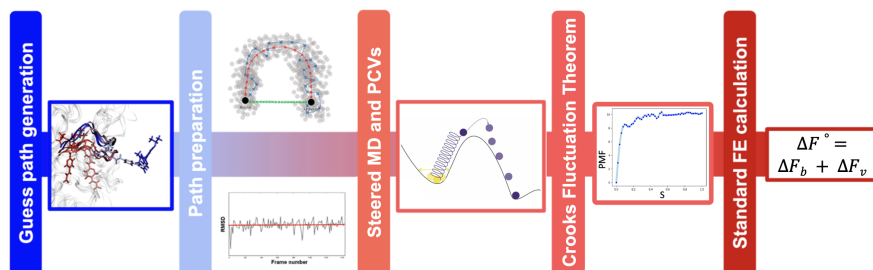


Figure 4.2: Graphical overview of the semi-automatic computational workflow

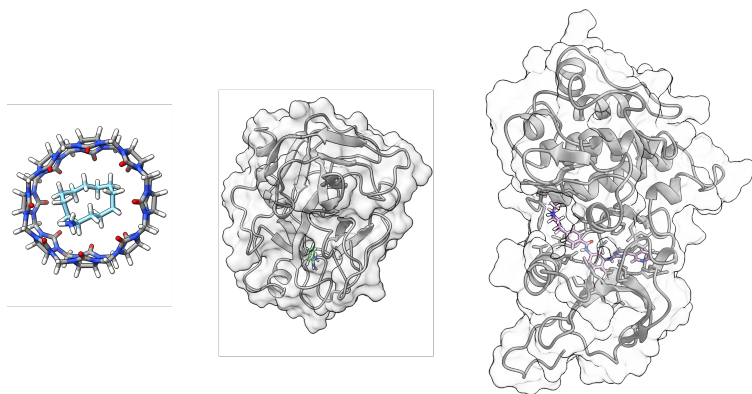


Figure 4.3: The three selected systems, arranged in order of increasing size and complexity. From left to right: HST-GST (curcubit[8]uril(CB8)-G8) serves as a toy model for fine-tuning and validating the pipeline; Trypsin-Benzamidine is used as a standard benchmark system for free energy methodologies; and Abl tyrosine kinase-Gleevec represents a challenging yet therapeutically significant system.

4.4 Results and Discussion

The proposed computational strategy is first tested on a simplified toy model to evaluate its validity in reconstructing the free energy profile along $S(x)$. After successful validation, the approach is applied to a more complex system: the Trypsin-Benzamidine binary complex. Finally, the method is used to study a challenging system involving the c-Abl tyrosine kinase and its inhibitor, Gleevec. The three chosen systems exhibit progressively larger dimensions and greater complexity, as illustrated in Figure 4.3. All simulations are conducted using GROMACS 2021.4 [204] integrated with PLUMED 2.9 [205].

4.4.1 CB8-G8 System

The host-guest complex formed by cucurbit[8]uril (CB8) and the guest molecule G8 serves as a simplified molecular model for validation purposes (depicted as the first molecule in Figure 4.3). This system was previously studied using Well-Tempered MetaD with PCVs to reconstruct its PMF [180]. The primary objective of this application is to evaluate the ability of our new protocol, which leverages bidirectional SMD and nonequilibrium estimators, in accurately reproducing the PMF along the defined PCVs.

The CB8-G8 complex is parameterized using the GAFF with AM1-BCC point charges. The system is solvated with TIP3P water molecules, and sodium and chloride ions are added to neutralize it. By using the previously defined reference path [180], we run 50 replicas of bidirectional SMD with PCVs in the NVT ensemble, using different steering velocities. Non-bonded interactions are handled with a 12 Å cutoff, and long-range electrostatic interactions are computed using the Particle Mesh Ewald method with a grid spacing of 1.6 Å. For production runs, all simulations start from the same initial configuration but with different initial velocities (i.e., utilizing different random number seeds). Additionally, to ensure accurate and reproducible results, PME cutoff adjustments by GROMACS are disabled during the calculations.

To evaluate the impact of pulling speed on the free energy profile along $S(x)$ and

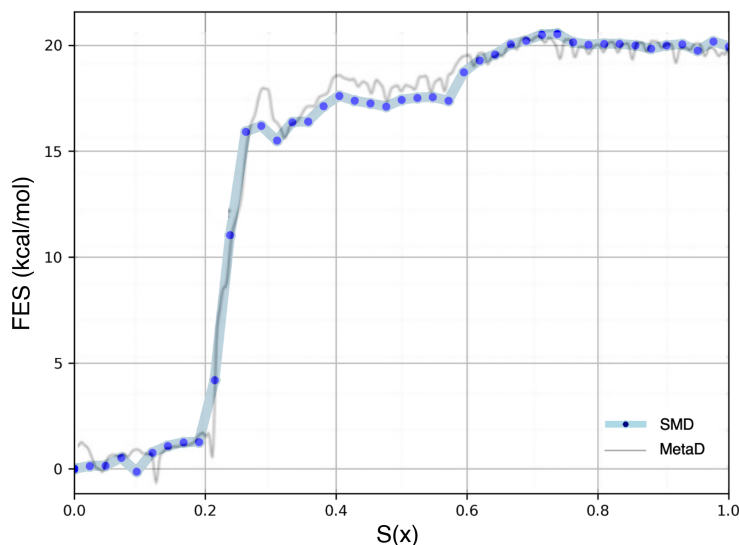


Figure 4.4: PMF of the CB8-G8 complex along $S(x)$ obtained by applying the CFT to 150 ns long SMDs on PCVs overimposed to PMF resulting from Ref. [180]. The overlay demonstrates nearly perfect quantitative consistency between the two results

the convergence of the estimator, SMD simulations are performed at five simulation times: 10 ns, 25 ns, 50 ns, 100 ns, and 150 ns. Then, all the PMFs (at different velocities) are reconstructed by using iteratively both CFT and JE. As demonstrated in Figure 4.4, the CFT-derived PMF from the longest simulations (i.e., 150 ns) show excellent agreement with previous MetaD results [180],

4.4.2 β -Trypsin-Benzamidine System

As a more realistic benchmark, we select the β -Trypsin-Benzamidine (TRY-BEN) binary complex, a widely studied model for testing computational approaches for protein-ligand binding [200, 128, 206]. In this system (depicted as the second molecule in Figure 4.3), the ligand-receptor interactions are mainly electrostatic forces and hydrogen bonding, giving an experimental binding free energy of -6.2 kcal/mol [207].

Following Ref. [208], the crystallographic structure with PDB code 3PTB is used as a starting point to model the TRY-BEN complex. The Trypsin protein is modeled using the AMBER99SB [71] FF, while the Benzamidine ligand is represented with GAFF parameters and AM1-BCC point charges. The system is solvated in

explicit TIP3P water molecules within a cubic box, with the protein placed 1.2 nm away from the box edges. Sodium and chloride ions are added to achieve a physiological salt concentration of 0.15 M and ensure electric neutrality. Then, the solvated and neutralized complex is energetically minimized using 50000 steps of the Steepest Descendent algorithm. This energy minimisation process is followed by three equilibration stages:

1. An initial *NVT* simulation of 500 ps. The Bussi-Donadio-Parrinello thermostat [87] is used to thermostat the system at 300 K, with harmonic restraints of 1000 kJ/mol applied to the heavy atoms of both the protein and ligand.
2. *NPT* simulation of 500 ps. The Berendsen Barostat [86] controls the pressure at 1 atmosphere, while the system is thermostated at 300 K. 1000 kJ/mol harmonic restraints are applied to the non-hydrogen atoms of both the protein and ligand.
3. *NPT* 1 ns equilibration simulation at 1 atmosphere and 300 K (using the same barostat and thermostat) where all atoms can move freely.

The structure obtained from the final equilibration step is used as the starting point for the production simulations in the *NVT* ensemble.

Path Collective Variables and nonequilibrium simulations

The initial guess path is generated by performing several 10 ns ABMD simulations in the *NVT* ensemble using BiKi [202], with the electrostatics-like CV [198] as described earlier. Chemical bonds involving hydrogen atoms are constrained by using the LINCS algorithm, enabling a 2 fs integration step. Non-bonded interactions are handled with a 12 Å cutoff, while long-range electrostatic interactions are computed using the PME method with a grid spacing of 1.6 Å. After identifying the most likely unbinding trajectory, path algorithms are employed to define a minimum free energy path, consisting solely of the ligand atoms. This optimized minimum free energy path is composed of 52 equidistant conformations representing the ligand (un)binding process.

Following this path, we perform bidirectional SMD simulations via PCVs with

different steering velocities in the NVT ensemble. The same parameters and SMD protocol used for the toy model are employed for this system as well. In theory, conducting fast SMD simulations with a small number of trajectories could enable quick free energy calculations. However, in practice, this approach may result in poor convergence of the estimates. To assess the accuracy of our method when using a finite number of nonequilibrium trajectories and high pulling speeds, we conduct 50 binding and 50 unbinding SMD simulations, varying the simulation durations from 10 to 150 ns.

From these bidirectional SMD simulations, we calculate the W_J performed on the system during the irreversible processes. These work profiles for the two limiting cases are illustrated in Figure 4.5. As shown in the upper section of Figure 4.5, during unbinding simulations the cumulative W_J starts close to zero, as the complex is initially in the bound state. As the simulation progresses, the cumulative W_J gradually increases, reaching eventually a plateau when the system is in the unbound state. Notably, for longer simulation times, the cumulative W_J tends to decrease, as slower steering velocities lead to less dissipated work and, consequently, lower W_J . In contrast, during binding simulations (lower section of Figure 4.5), W_J begins close to zero, since the ligand is initially in the unbound state within the solvent. As the simulation progresses, the absolute value of the work increases but becomes more negative (to maintain consistency with the unbinding simulations). For very fast SMD simulations, the total W_J is largely positive, reflecting a higher amount of dissipated work. However, for longer simulations, the dissipated work is reduced, and negative values of W_J become more likely.

PMF and Binding Free Energy estimation

With our semi-automatic algorithm, both the CFT and JE derived PMFs at different simulation times along $S(x)$ are reconstructed, as shown in Figure 4.6. Here, $S(x) = 0$ corresponds to the bound state, while $S(x) = 1$ to the unbound state (normalized $[1, p]$ to $[0, 1]$). As can be seen, the longer the simulation times, the lower the free energy values associated with the fully solvated ligand state (represented by the plateau in the PMF).

To compare the free energy profiles obtained using the bidirectional CFT estimator and the unidirectional JE, we reconstruct the PMFs for both binding and unbinding simulations with the JE. In unbinding simulations (Figure 4.6b), the trend seen in the CFT-derived PMFs for different simulation times is also observed in the Jarzynski-based PMFs. However, the plateaus corresponding to the unbound state occur at higher free energy values when using this estimator. In binding simulations, applying JE results in significant noise in the PMF (Figure 4.6c), especially at higher pulling speeds. This is due to the high dissipated work, which influences the exponential average used in the Jarzynski-based PMF calculation. Despite the noisy PMF from the 10 ns SMD, the trend in the binding simulations reverses compared to the CFT-derived profiles. Longer binding simulations generate more negative values of Jarzynski work, allowing more replicas to reach negative values in the bound state.

Subsequently, we compute the binding free energy from the obtained PMFs. Binding free energy estimation requires the calculation of the ratio between the bound and unbound partition functions. This ratio is computed as the PMF integral over the bound region divided by the integral over the unbound region (Eq. 4.5). The value of $S(x)$ discriminating these two regions, as suggested by Doudou *et al* [200], corresponds to the point where the PMF stabilizes within the statistical noise, and the ligand is solvated. Accordingly, a discriminating configuration between the 6th (partially docked conformation) and the 8th (partially solvated state) points of the free energy profile results meaningful. In this study, the discriminating frame is selected corresponding to the 8th configuration.

Lastly, all the PMF-derived binding free energies are corrected with the standard volume correction term, resulting in ΔF° comparable with the experimental binding affinity. Outcomes for all simulation times are reported in Table 4.1 and summarized in Figure 4.7. These results indicate that, as the simulation time increases, the estimated binding free energies progressively converge toward the experimental values. Particularly, for 100 ns SMD simulations, the binding free energy is -6.6 ± 0.5 kcal/mol, demonstrating a strong agreement with the experimental value. Notably, even with longer simulation times (i.e., 150 ns), we do not observe substantial

changes in the estimate, reaching a binding free energy of -6.5 ± 0.4 kcal/mol.

Moreover, to further refine our estimates, we extrapolate the free energy value by evaluating the complete free energy series as a function of steering time. The binding free energy convergence is modeled using a parameterized exponential function:

$$y = a(d - \exp(-bx + c)) \quad (4.6)$$

where y is ΔF , x is the simulation time and a, b, c , and d are the exploration's parameters. For a quasi-static transformation (at infinite time), the extrapolated binding free energy is -6.5 kcal/mol, showing a good agreement with the experimental value.

The comparison between the bidirectional CFT and the unidirectional JE reveals distinct biases in the latter. Specifically, applying the JE to binding simulations tends to overestimate the binding free energy, whereas its application to unbinding simulations leads to an underestimation. Consequently, the JE estimates from binding simulations serve as the upper bound for the CFT-derived estimates, while those from unbinding simulations represent the lower bound as represented in Figure 4.7 (it is a common characteristic of such estimation methods [149]). In accordance with the second law of thermodynamics, the Jarzynski work is always equal or greater than the free energy difference, thus $\langle W_J^f \rangle$ and $\langle -W_J^b \rangle$ represent the limiting bounds of ΔF . The free energy difference obtained by integrating forward and backward transformations ($\Delta \hat{F}_{CFT}$) is expected to be constrained by the forward and backward Jarzynski estimates ($\Delta \hat{F}_f$ and $\Delta \hat{F}_b$) following the relation: $\Delta \hat{F}_f \leq \Delta \hat{F}_{CFT} \leq \Delta \hat{F}_b$. However, this condition is not strictly required [149]. Nonetheless, the use of one-sided free energy differences as bounds is generally valid, as the average work serves as the first cumulant approximation of the JE.

Binding simulations show no clear trend in free energy or error reduction, but at 150 ns, the estimate (-5 ± 1 kcal/mol) aligns fairly well with experiments. In contrast, unbinding simulations exhibit slow convergence, with the 150 ns result (-10.2 ± 0.3 kcal/mol) still deviating from experimental values. Therefore, for the β -Trypsin-Benzamidine complex, our study highlights the robustness and supe-

	10 ns	25 ns	50 ns	100 ns	150 ns
JE binding	-4 ± 2	-2.7 ± 0.9	-3.7 ± 0.8	-5 ± 1	-5 ± 1
JE unbinding	-20 ± 1	-17.1 ± 0.3	-14.2 ± 0.4	-10.6 ± 0.7	-10.3 ± 0.3
CFT	-12 ± 1	-8.9 ± 0.3	-8.1 ± 0.3	-6.6 ± 0.5	-6.5 ± 0.4

Table 4.1: Standard binding free energies (kcal/mol) for TRY-BEN estimated using the JE and CFT for increasing simulation times.

riority of the two-sided CFT estimator over the one-sided JE. The JE struggles with poor convergence due to its reliance on rare low-work events, which dominate the exponential average. These rare trajectories become even scarcer as the SMD pulling speed increases, significantly amplifying the computational effort needed for convergence. Consequently, the number of simulations required to achieve reliable results with the JE can become impractically large. The bidirectional CFT estimator, which leverages both forward and backward simulations, may appear to involve higher computational costs compared to unidirectional estimators due to the need to simulate both binding and unbinding processes. However, as highlighted by Pohorille *et al.* [139], and supported by our results, two-sided estimators are more reliable for estimating free energy estimates within the same computational time, demonstrating their efficiency and robustness over one-sided approaches.

In addition, we evaluate the sensitivity of the binding free energy estimates to the discriminating configuration selected to compute the ratio of partition functions of the bound and unbound regions. To assess the impact of this PCV $S(x)$ cutoff selection on the estimation, we perform a sensitivity analysis: we extend the range of the PCV $S(x)$ between the 4th and the 16th configurations, which corresponds to the fully unbound state. The analysis of the 150 ns simulations (Figure 4.8) shows that variations in the binding free energy across different discriminating configurations (x^*) remain minimal, within 1 kcal/mol for all estimators, even when extending to the 16th configuration for the CFT estimator. This analysis indicates that our estimates are insensitive to the choice of x^* , confirming the robustness of using the ratio of partition functions for computing binding free energies.

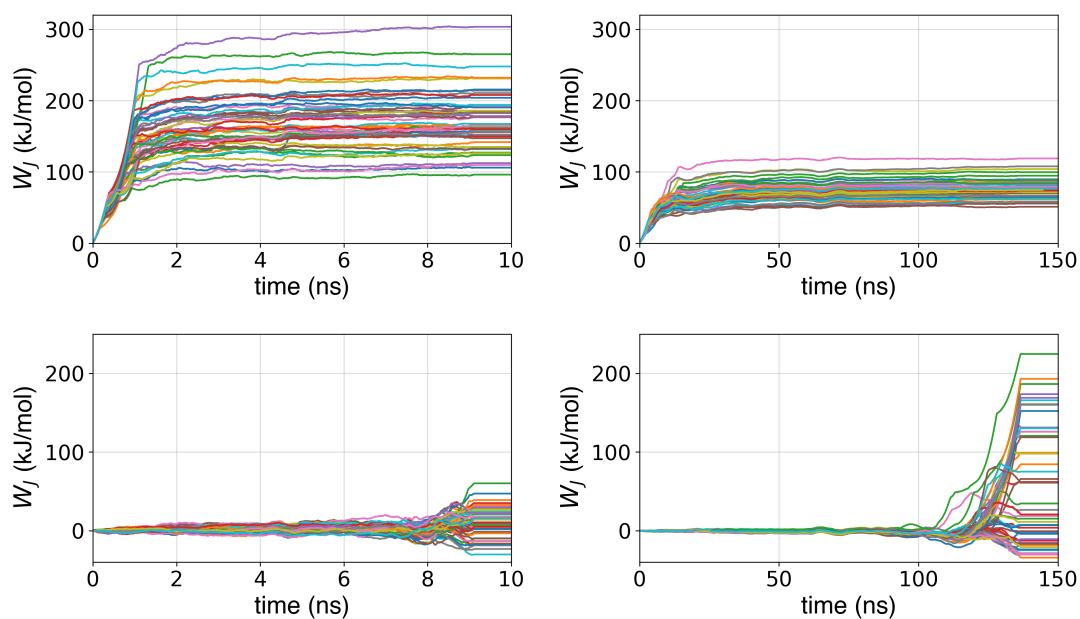


Figure 4.5: W_J profiles for the TRY-BEN system are presented for both unbinding (top) and binding (bottom) simulations, with measurements taken over simulation durations of 10 ns and 150 ns.

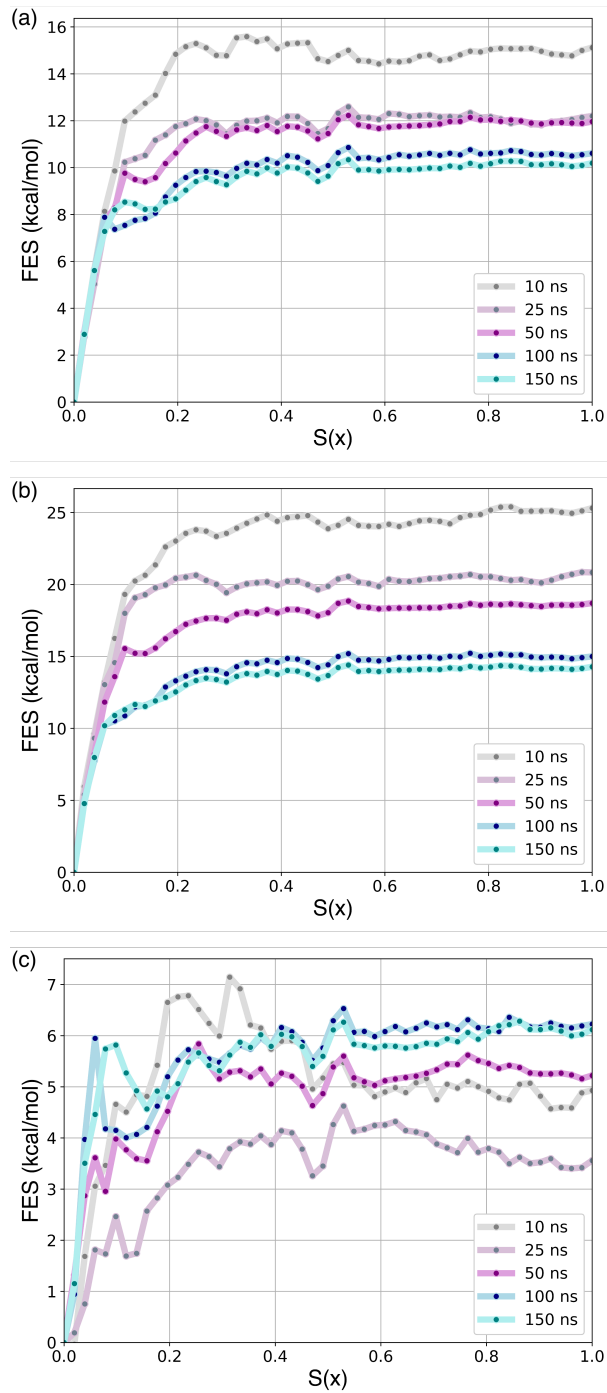


Figure 4.6: Free energy profiles along $S(x)$ for TRY-BEN obtained by applying the (a) CFT to binding and unbinding SMD, (b) JE to unbinding SMD, and (c) JE to binding SMD for increasing simulation times.

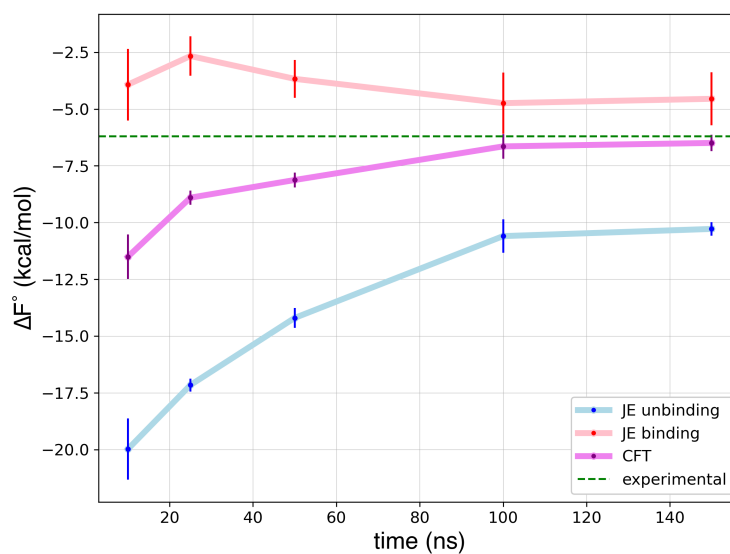


Figure 4.7: Standard binding free energies (kcal/mol) for TRY-BEN estimated using the JE and CFT for increasing simulation times compared to the experimental value of -6.2 kcal/mol. Uncertainty is three times the bootstrap standard deviation.

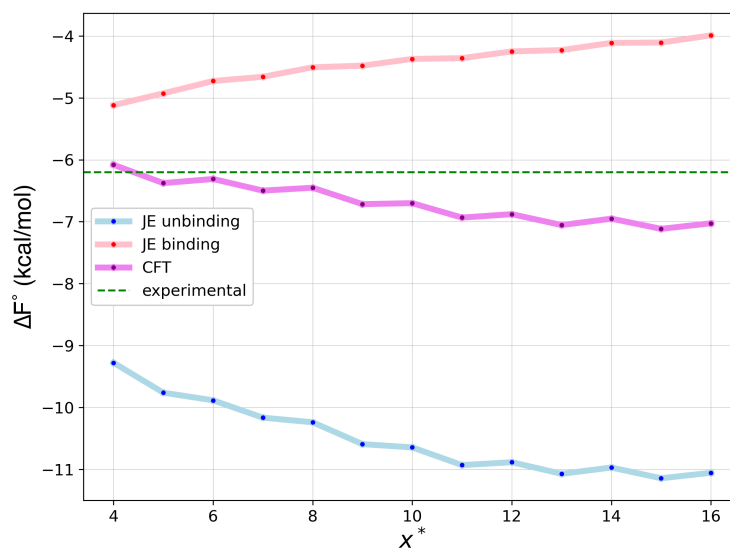


Figure 4.8: Standard binding free energies (kcal/mol) variation for TRY-BEN computed using the JE and CFT for 150 ns SMD simulations as a function of the discriminating configuration compared to the experimental value of -6.2 kcal/mol.

Convergence analysis

For the TRY-BEN system, we also perform a convergence evaluation of the binding free energy with the number of replicas. As previously explained, to estimate the binding free energy, 50 binding and unbinding simulations are performed at different simulation times. Despite the inherent bias and variance errors in free energy estimators due to finite samples (i.e., finite number of simulations), 50 replicas result sufficient to achieve convergence. Moreover, the variance dominates over bias in all the estimates.

Using bootstrapping resampling with 500 repetitions for each number of replicas, we conduct an extensive statistical analysis (with the error defined as 3 times the standard deviation), showing that with 150 ns SMD simulations, the CFT estimate stabilizes within 1% of the final value using around 20 replicas. Errors at this point are less than 1 kcal/mol, representing under 5% of the binding free energy estimate (Figure 4.9a).

In contrast, achieving convergence with the one-sided JE is more difficult due to its poor performance with limited simulations. Additionally, its biased nature and the kinetic asymmetry between binding and unbinding mean that convergence in forward simulations does not guarantee convergence in backward simulations. Figure 4.9b illustrates the trend of the binding free energy estimate and its relative error as the number of replicas increases for unbinding simulations, while Figure 4.9c shows the same for binding simulations. For Jarzynski unbinding estimates, convergence is reached within fewer than 10 simulations of 150 ns, with a bootstrap error below 2%. However, despite low variance, the estimates stabilize around -10 kcal/mol, deviating from the experimental value of -6.2 kcal/mol. This indicates that while the number of replicas is sufficient for low variance, true convergence requires longer simulations to observe a stable reduction in dissipated work. In contrast, Jarzynski binding estimates require more replicas for convergence, with stability achieved around 40 replicas for 150 ns simulations, yet still showing a percentage error exceeding 10% even with 50 replicas. The estimates resulting from at least 40 binding replicas are about -5 kcal/mol, thus less biased than the unbinding estimates. This can be attributed to the significant inherent variability

of the binding process. In fact, the binding process introduces high variance but maintains limited bias in JE due to relatively low kinetic barriers and dissipated work. In contrast, the unbinding process, characterized by higher dissipated work, exhibits significant bias but reduced bootstrap variance in the JE due to its more constrained nature. Remarkably, the CFT estimator is able to get the best of these two scenarios.

Considering the convergence analysis of the estimates with respect to the number of replicas for simulations at other time scales, both CFT and JE estimates at 100 ns exhibit similar behavior to those at 150 ns, confirming the reliability of CFT for the TRY-BEN system. However, achieving convergence in faster simulations is more challenging, leading to greater errors compared to the experimental value.

In conclusion, our novel nonequilibrium approach using the CFT estimator achieves convergence with 40 simulations (20 binding and 20 unbinding) of 100 ns each, totaling 4 μ s of simulation time. While this may seem less efficient than other methods, the ability to run all simulations in parallel allows convergence to be reached in just 100 ns of real-time execution when utilizing high-performance computing resources. A deeper explanation of this statistical analysis is available in Ref. [P1].

4.4.3 Abl-Gleevec System

The c-Abl tyrosine kinase-Gleevec complex (shown as the third structure in Figure 4.3) is a highly relevant system, due to the therapeutic role of kinases in treating various diseases and the strong inhibitory effect of Gleevec on c-Abl ($\Delta F_b = -10.8$ kcal/mol [209]). To further challenge our protocol, we apply it to this pharmaceutically relevant complex. Gleevec, a widely used anticancer drug today, exhibits high molecular flexibility and a large size, thereby introducing additional complexity and making this system more demanding than previously studied benchmarks. The protocol presented for TRY-BEN system is applied to Abl-Gleevec, generating a putative minimum free energy path with 90 molecular configurations describing the ligand (un)binding process. Then, two-sided SMD simulations are run at three simulation times: 10 ns, 50 ns, and 100ns. Notably, the large size of the solvated Abl-Gleevec system, with hundreds of thousands of atoms, significantly increases

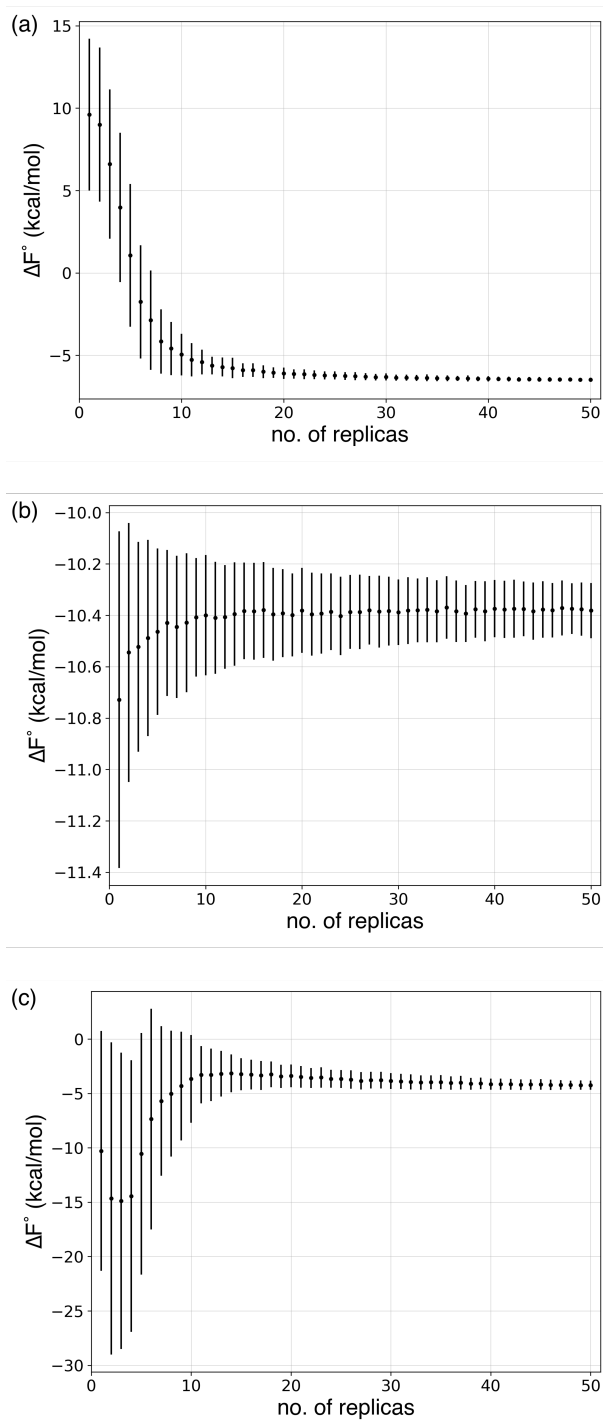


Figure 4.9: Standard binding free energies (kcal/mol) for TRY-BEN estimated using the (a) CFT and JE on (b) unbinding and (c) binding simulations for 150 ns SMD as a function of the number of replicas. Uncertainty is three times the bootstrap standard deviation.

the computational cost of all the simulations.

As discussed earlier, Jarzynski work profiles for binding and unbinding simulations are reconstructed at each simulation time (Figure 4.10). In the early stages of unbinding simulations, the W_J remains low, corresponding to the bound state. As the unbinding progresses, W_J gradually increases, eventually reaching a plateau at the unbound state. Similar to the results observed for TRY-BEN, this plateau occurs at lower W_J values as the simulation time increases, primarily due to the dissipation reduction. For binding simulations, W_J starts near zero while Gleevec is in the bulk solvent. As the simulation progresses and the binding event occurs, the work increases rapidly, eventually reaching values comparable to or even higher than those observed during the unbinding simulations. Interestingly, the observed behavior contradicts the expected trend. Binding appears to be an unfavorable event, even for the 100 ns simulations, where only a small reduction in the cumulative W_J is observed. This suggests that for this complex system, 100 ns SMD still corresponds to a high pulling speed, resulting in significant dissipated work. Due to this high dissipated work, nonequilibrium estimators fail to produce converged estimates, unlike in the TRY-BEN case, where the pulling time selected was sufficient. The binding free energy estimates from 50 binding and 50 unbinding 100 ns simulations using the unidirectional JE are -72 kcal/mol for unbinding and +57 kcal/mol for binding. In contrast, the bidirectional CFT estimator gives a value of -4 kcal/mol. While the bidirectional CFT estimator reduces the large discrepancies between the forward and backward processes compared to the experimental value, the result still deviates from the experimental data. The associated estimated error is ± 8 kcal/mol, indicating that the simulations have not yet converged. Additionally, the trend of the binding free energy estimate for Abl-Gleevec as a function of simulation time shows no signs of flattening by 100 ns. In comparison, for TRY-BEN, the trend's slope decreases with longer simulation times, converging at 100 ns (Figure 4.7).

The primary factor contributing to the lack of convergence is the significant dissipation observed during binding and unbinding events. Although reducing the pulling speed in SMD simulations could mitigate this effect, doing so would demand sim-

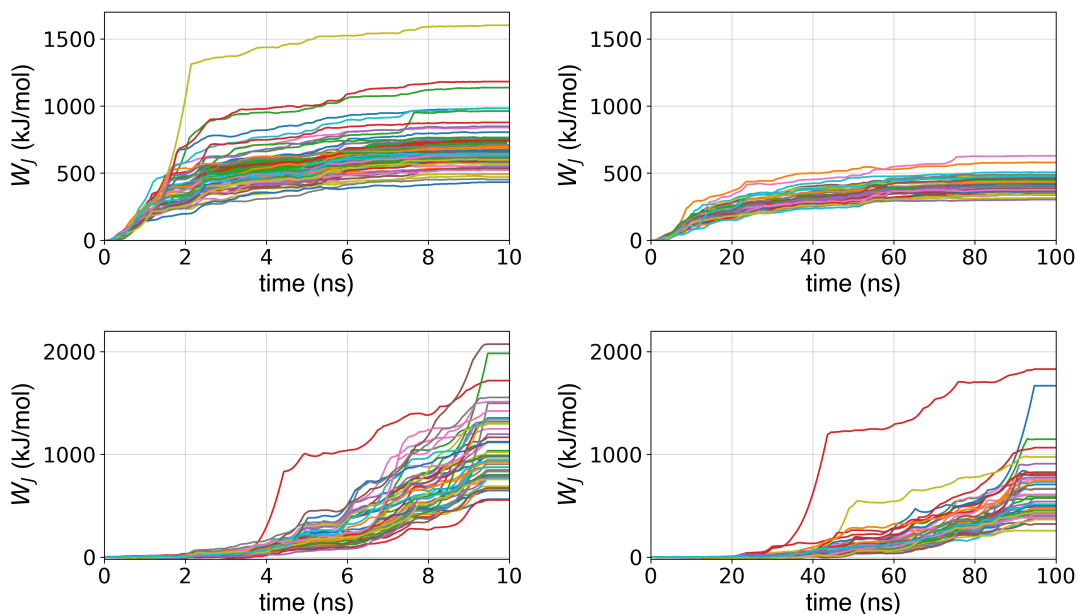


Figure 4.10: The Jarzynski work profiles are presented for both unbinding (upper section) and binding simulations (lower section) for Abl-Gleevec system. Measurements are taken over simulation time of 10 ns and 100 ns.

ulation times exceeding 100 ns per trajectory, rendering the approach impractical for drug discovery applications. Consequently, alternative strategies must be considered when dealing with large and complex systems. We believe these should focus on minimizing dissipation and involve the development of specific criteria for optimally defining the reference path and PCVs. The Abl-Gleevec system, in particular, highlights the current limitations of our protocol, which will be discussed and addressed in detail in the following sections.

4.5 Conclusion

Starting from previous research on binding free energy calculations [180], we adapt this methodology to a nonequilibrium framework. The proposed approach is first validated on a simplified toy model before being applied to the well-established trypsin-benzamidine system as a benchmark. Finally, we test the new computational pipeline on a more complex pharmaceutical system to further evaluate its

performance. The results confirm the applicability of this innovative approach while also identifying its limitations. Furthermore, the study underscores the relevance and effectiveness of Jarzynski [144, 145] and Crooks [148, 186] estimators in the context of nonequilibrium free energy calculations for biological systems.

In particular, this study explores the integration of nonequilibrium methods with PCVs, leveraging their computational advantage of natural parallelism. The results demonstrate that this approach can yield accurate binding free energy estimates for small to moderately sized systems while providing detailed molecular-level insights. Moreover, our investigation highlights that, while the W_J is statistically inefficient when applied via the JE, combining W_J with the CFT produces reliable results. This method aligns with previous findings [210] and emphasizes the importance of distinguishing the inefficiency of the JE.

Therefore, for moderate-sized systems, our CFT-based results correlate well with experimental data. However, the method encounters challenges with larger, more complex systems, such as the Abl-Gleevec complex, where results are less accurate. Efforts to refine the protocol for its application to larger systems while maintaining reasonable computational costs are presented in the following sections.

In conclusion, the combined use of CFT, SMD via PCVs, and W_J is a promising approach for estimating binding free energy in systems of moderate size and complexity. This method shows potential for broader applications, such as protein-protein interactions. Furthermore, its suitability for massively parallel computational architectures and straightforward convergence assessment makes it a valuable tool for computational drug discovery researches.

Chapter 5

Minimizing dissipation in nonequilibrium binding free energy simulations [P2]

5.1 Aim and Objectives

As anticipated, equilibrium methods for free energy estimation are now well established, whereas nonequilibrium approaches offer both promising computational advantages and inherent challenges. While nonequilibrium simulations enable straightforward parallelization, the substantial irreversible work produced during the steering process can impede the convergence of free energy estimators. Starting from our established nonequilibrium pipeline and recognizing its limitations (Section 4 [P1]), this work aims to investigate key factors that contribute to the dissipation observed during irreversible SMD simulations. Our focus lies in examining the influence of both the water model and the parametrization of PCVs on the convergence rate of nonequilibrium free energy estimators. In the end, we want to propose practical strategies to accelerate convergence and reduce dissipation in protein-ligand binding free energy computation using physical path-based approaches.

5.2 Introduction

In the context of MD simulations, several computational approaches have been developed to estimate the free energy difference between two equilibrium states of a system. Among them, nonequilibrium techniques have gained increasing attention in recent decades [105]. These methods rely on a crucial measured quantity, the Jarzynski work (Eq. 2.60), which represents the cumulative time-dependent change in the system's Hamiltonian [143]. According to the second law of thermodynamics, in a quasi-static ideal transformation, W_J corresponds to the difference in free energy. However, for nonequilibrium, irreversible transformations in real world systems, the work typically exceeds the free energy difference by an additional amount known as the dissipated work, W_J^{diss} (Eq. 2.61) [144].

During a single simulation, the observed Jarzynski work is a random variable that can be decomposed into two contributions: its deterministic part, which is the free energy difference, and the dissipation that occurs during the process. Therefore, in addition to Eq. 2.61, the W_J^{diss} can be formulated as:

$$W_J^{diss} = \langle \Delta F_{AB} + W_J^{diss, traj} \rangle - \Delta F_{AB} \geq 0 \quad (5.1)$$

where $\Delta F_{AB} = F(B) - F(A)$, and $W_J^{diss, traj}$ is the incurred dissipation.

In the nonequilibrium scenario, calculating ΔF_{AB} requires nonequilibrium estimators in order to consider W_J values obtained from multiple irreversible trajectories (executed under the same simulation protocol). As highlighted in Chapter 2.5, the primary nonequilibrium estimators include JE [144, 145] and CFT [148, 186], which follow different estimation approaches. Particularly, JE determines the equilibrium free energy difference between the system's states by relating it to the external W_J through an exponential averaging method (Eq. 2.62). This exponential average is computed using N sampled unidirectional trajectories, and due to its inherent properties, JE is classified as a unidirectional estimator. Unidirectional JE is commonly applied for free energy estimations, but it is known to present challenges in statistical convergence (a limitation we have also demonstrated in our previous

work [P1]). This issue arises because JE estimator relies on an exponential average, where the lowest W_J values (often infrequently sampled) heavily influence the result. As a consequence, a significantly large number of finite-time simulations may be required to obtain reliable convergence.

To overcome this limitation and enhance the accuracy of free energy estimates, the bidirectional nonequilibrium CFT estimator is a valuable option (Eq. 2.63). Like the unidirectional JE, CFT is also affected by dissipation. However, its bidirectional framework yields improved efficiency in free energy estimation. By leveraging CFT, the computation of ΔF_{AB} is obtained by determining the point where the W_J forward and backward distributions intersect. In contrast, JE requires precise sampling of the tails of either the forward or backward W_J distribution. Consequently, the bidirectional nature of the CFT facilitates a more straightforward discounting of dissipated work during the estimation process. As discussed in the literature [139] and reinforced by our previous work [P1], CFT is therefore more reliable than JE for estimating free energy differences within a given computational budget.

In the last decade, both CFT and JE have been utilized in free energy calculations, whether following alchemical routes [146, 152] or path-based strategies [147]. However, convergence challenges frequently hinder the practical use of these estimators. Their ability to reliably approach the correct free energy estimate is heavily influenced by the amount of dissipated work produced during nonequilibrium simulations. Reducing this dissipated work leads to lower W_J , which accelerates the convergence of nonequilibrium estimators to the true free energy difference. Thus, minimizing dissipation becomes crucial for improving convergence and developing efficient nonequilibrium strategies. Dissipation depends not only on the transformation path but also on the system’s physicochemical parameters, including the endpoint Hamiltonians. All these factors must be considered when designing novel nonequilibrium protocols.

As previously noted, nonequilibrium methods are computationally attractive due to their inherent parallelizability, making them particularly suitable for resource-intensive simulations such as protein–ligand binding studies. Consequently, they serve as valuable tools for estimating binding free energies, ΔF_b (Eq. 2.52), in

the context of drug discovery. In our initial work (Chapter 4), we investigated the combination of nonequilibrium SMD via PCVs and nonequilibrium estimators, proving that CFT offers greater robustness and is less sensitive to convergence issues than JE [P1]. Moreover, the approximate number of samples required for the convergence of these nonequilibrium estimators is known to be influenced by the amount of dissipated work during the transformation, following an exponential relationship [139]. As dissipation increases, the number of trajectories needed for convergence of ΔF_b becomes rapidly impractical. Therefore, the effectiveness of our nonequilibrium protocol is closely tied to its ability to minimize dissipated work throughout the process.

Based on our original nonequilibrium pipeline [P1], here we deeply examine the challenges of estimating protein-ligand binding free energy using physical path-based approaches. We identify two key factors that contribute to dissipation in irreversible SMD simulations: the choice of the water model and the parametrization of PCVs. We want to define general applicable suggestions for selecting an appropriate water model in nonequilibrium scenario and provide protocol-specific insights for PCVs definition. Overall, we believe that strategies proposed in this study significantly enhance both the convergence and accuracy of nonequilibrium free energy estimations.

5.2.1 Water Models

In protein-ligand binding simulations, accurately representing water behavior is crucial for obtaining meaningful results from MD simulations [91, 92]. Indeed, more than 90% of the solvated system consists of the surrounding aqueous environment, which plays a crucial role in contributing to both the potential and kinetic energy of the system. As explained in Chapter 2.3.6, explicit water models are typically preferred for this purpose, as they offer sufficient accuracy for solvent modeling. However, no single explicit model can perfectly replicate all experimental properties of water [96]. As a result, selecting the appropriate solvent model requires careful consideration of its ability to capture the specific water-related characteristics that are most relevant to the biological process under investigation

[93, 94]. In protein–ligand binding simulations, fixed-charge 3- and 4-point models remain the most widely used, striking a balance between accuracy and computational efficiency, despite the availability of more advanced models.

In this study, we systematically investigate how different water models affect dissipation, the reconstruction of the potential of mean force in nonequilibrium simulations, and the estimation of binding free energies. In contrast with equilibrium methods, nonequilibrium approaches like SMD are highly sensitive to kinetics and energy barriers, influencing the amount of dissipated work. In this scenario, the kinetic and mechanical properties of the water model play a crucial role, particularly in small protein-ligand complexes with limited degrees of freedom. For these systems water results the primary contributor to dissipation. Our analysis focuses on assessing the effect of widely used water models on work dissipation and their suitability for nonequilibrium simulations. For this work, we select two protein-ligand complexes: β -Trypsin-Benzamidine, and glycogen synthase kinase 3 beta (GSK-3 β) with its ATP-competitive inhibitor.

5.2.2 Path collective variables

During nonequilibrium SMD simulations, the system evolves along a pathway connecting the initial and final states, which necessitates the definition of appropriate CVs. CVs are function of a few coordinates that track the process progression, and by using nonequilibrium estimators, the PMF can be reconstructed. To meet the bidirectional simulation requirements of the CFT, our pipeline utilizes a specific type of CVs, namely the PCVs [115]. PCVs ($S(x)$ and $Z(x)$ formulated in Eq. 2.41 and 2.42, respectively) are particularly suitable as they represent the position of a point in the configurational space relative to a predefined path, enabling precise tracking of both forward and backward transformations. Once the optimal path describing the unbinding event is identified, the free energy profile along the PCV $S(x)$ can be reconstructed, leading to the correct estimation of binding free energy. However, if the reference path includes a mechanism that deviates from the physically true binding process, it may introduce significantly high energy barriers, resulting in increased dissipated work. This dissipation can hinder the accurate

estimation of the binding free energy.

In this study, we examine the impact of the putative minimum free energy path and PCVs on both the reconstructed free energy profile along $S(x)$ and the estimation of free energy from nonequilibrium simulations. We specifically investigate the T4 Lysozyme-N-phenylglycinonitrile complex, which presents unique challenges due to the ligand being deeply embedded within the protein [211], a scenario not seen in the systems studied before [P1]. Additionally, we provide strategies for selecting an appropriate reference path, taking into account the specific characteristics of the system under consideration.

5.3 Methods

To run SMD simulations with PCVs, reconstruct the PMF along $S(x)$, and estimate the binding free energy, we utilize our validated computational pipeline [P1]. Our approach is based on the bidirectional CFT, which offers improved robustness and faster convergence compared to the JE. However, in this study, both estimators are applied to examine the varying amounts of dissipated work generated during the binding and unbinding simulations.

The five key steps of our nonequilibrium approach have been previously outlined in Section 4 and summarized in Figure 4.2. Accordingly, multiple SMD simulations for both binding and unbinding processes are conducted iteratively until the binding free energy estimate converges. The computational performance for the analyzed systems ranged between 80 and 120 ns per day, utilizing a dual-socket node equipped with Intel(R) Xeon(R) Silver 4210 CPUs and two Tesla V100 GPUs.

5.3.1 Selected systems

To evaluate the influence of water models on the binding free energy estimate and the PMFs, we analyze two different complexes: β -Trypsin-Benzamidine binary complex (TRY-BEN), and the glycogen synthase kinase-3 beta (GSK-3 β) complex. We start our systematic study with the TRY-BEN complex, which has an experimental binding free energy of -6.2 kcal/mol [207]. Based on literature [211] and

our findings, Benzamidine is partially buried within the β -Trypsin pocket. This analysis includes both three-point water models (TIP3P, TIP3P-flexible, OPC3) and four-point models (TIP4P, TIP4P-flexible, OPC). Since we have previously examined TRY-BEN using TIP3P combined with SMD, PCVs, and CFT, we use the reference path in TIP3P from our earlier work [P1]. TRY-BEN complex is solvated with all the selected water models, resulting in 12847 and 12777 water molecules for three- and four-point models, respectively. Each system undergo the same minimization and equilibration steps as those performed for TIP3P.

To further evaluate the influence of water models on binding free energy estimates, we analyze a second biologically relevant system: GSK-3 β and its ATP-competitive inhibitor 3-azanyl-N-(3-methoxypropyl)-6-[4-(4-methylpiperazin-1-yl)sulfonylphenyl]pyrazine-2-carboxamide, a potential therapeutic target for Alzheimer’s disease. Unlike TRY-BEN, this complex shows a shallow and exposed binding pocket, with an experimental binding free energy of -10.5 kcal/mol [212]. The system is first solvated in a cubic box using TIP3P water molecules, based on the crystallographic structure with PDB code 4ACH. The protein is represented using the AMBER14SB FF [63], while the ligand is modeled with GAFF [73] and AM1-BCC point charges, assigning a +1 charge to the nitrogen in the 4-methylpiperazine group (the detailed protocol is reported in Ref. [198]). To investigate the influence of various solvent models while also considering the TRY-BEN results, we conduct additional simulations employing the three-point OPC3 model and the four-point TIP4P and OPC models. Each system is solvated in a cubic box with a 1.2 nm buffer between the protein and box edges, resulting in 19213 water molecules for OPC3 and 19519 for TIP4P/OPC simulations. To ensure electric neutrality, Sodium and Chloride ions are added to the systems. The fully solvated protein-ligand complexes then undergo energy minimization and equilibration, following the same protocol used for TRY-BEN.

Finally, informed by the results on the most effective water model for minimizing dissipation, we evaluate how inaccuracies in the reference path affect binding free energy calculations. We explore how including different system degrees of freedom in the reference path, as well as the effect of the upper wall on $Z(x)$, influence work

dissipation and convergence in nonequilibrium simulations. For this purpose, we select T4 Lysozyme (Figure 5.15), a protein consisting of an N-terminal domain (residues 1–70) and a C-terminal domain (residues 71–162), which forms an eight α -helices barrel-shaped domain where the binding site is located. The binding cavity properties vary with mutations: the L99A mutant has an apolar, solvent-free pocket, whereas the L99A/M102Q mutant introduces a polar cavity able of hydrogen bonding [213]. While previous research has extensively analyzed ligand unbinding pathways in the L99A mutant, we focus on the less-studied L99A/M102Q double mutant. Here, the M102Q substitution creates a buried more polar binding site. Indeed, the replacement of Leucine with Alanine at position 99 generates a 150 Å³ hydrophobic pocket within the C-domain of the protein. However, the additional Methionine-to-Glutamine mutation at position 102 introduces a more polar engineered cavity, which remains shielded from the bulk solvent. Particularly, we select T4 Lysozyme L99A/M102Q in complex with N-phenylglycinonitrile (PDB ID: 2RBN), a challenging system due to the deeply buried position of the ligand [211, 12]. The binding interactions primarily involve electrostatic forces and hydrogen bonding (notably, a hydrogen bond between the (phenylamino)acetonitrile’s amino group and Gln102), with an experimental binding free energy of -5.52 ± 0.18 kcal/mol [207]. The complex is modeled using AMBER99SB [71] for the protein and GAFF with AM1-BCC charges for the ligand (with a total charge 0). Since the crystallographic PDB structure lacked several lysine side chains, Swiss Model [214] is used for preliminary modeling. Following the same protocol as for previous systems, the complex is solvated with TIP3P water, neutralized, and equilibrated. Additional information on the systems setup can be found in Ref.[P1, P2].

5.4 Results and Discussion

During nonequilibrium transformations, dissipation leads to irreversible processes where the average work exceeds the free energy difference. This excess work is defined as dissipated work (Eq. 2.61). Thus, the practical application of nonequilibrium estimators is frequently limited by convergence challenges, as their accuracy in obtaining reliable estimates is strongly influenced by the extent of dissipation

occurring during simulations. This can be seen by substituting the definition of W_J in the JE equation (Eq. 2.62) with the expression of the dissipated work (Eq. 2.61)

$$\begin{aligned}\exp(-\beta\Delta F_{AB}) &= \langle \exp(-\beta(W_J^{diss, traj} + \Delta F_{AB})) \rangle \\ 1 &= \langle \exp(-\beta W_J^{diss, traj}) \rangle\end{aligned}\tag{5.2}$$

Here, a zero dissipation would lead to a straightforward identity.

Therefore, reducing dissipation is crucial for optimizing nonequilibrium protocol. To achieve this, we systematically analyze two major contributors to dissipation: the choice of water model and the definition of PCVs. Our study aims to develop a nonequilibrium approach that effectively minimizes dissipation in both aspects.

5.4.1 Dissipation due to water model

In the initial part of the study, we examine how the choice of water model affects the dissipated work during irreversible simulations. Particularly, we quantify the dissipated Jarzynski work (Eq. 2.61) with the difference between the average final cumulative Jarzynski work (Eq. 2.60) obtained from SMD and the converged CFT estimate, $\Delta\hat{F}_{CFT}$. W_J^{diss} is evaluated for binding and unbinding simulations by changing the water model. For this systematic study, both the TRY-BEN and GSK-3 β complexes are analyzed.

Considering the TRY-BEN system, the reference path previously established with the TIP3P model is assumed to remain consistent across different water models. This assumption is valid as the path was derived from unbinding simulations, where water network displacement is not the key factor. Unlike binding simulations, which may exhibit slight variations depending on the water model, the unbinding process remains largely unaffected. Therefore, the same TIP3P-derived path is applied in all bidirectional SMD simulations.

For the GSK-3 β system, we determine the putative minimum free energy path using our protocol, resulting in an optimized reference path comprising 118 equidistant ligand conformations. This path is generated from ABMD unbinding simulations performed in a TIP3P-solvated system. This path captures the ligand unbinding

process and includes additional fictitious end-point conformations. Similar to TRY-BEN case, it is assumed that the path remains unchanged across different water models, allowing its use in all bidirectional SMD simulation replicas.

To evaluate the impact of water FFs and steering velocities on the binding free energy estimates, bidirectional SMD simulations are conducted at three different simulation times with different water models. Following the same simulation setup outlined previously [P1], 10 ns, 50 ns, and 100 ns bidirectional simulations are run. To achieve stable and converged estimates, the TRY-BEN system required 50 replicas for 10 ns, 30 replicas for 50 ns, and 20 replicas for 100 ns. Similarly, the GSK-3 β system required 30, 25, and 15 replicas for the 10 ns, 50 ns, and 100 ns simulations, respectively. This results in total simulation times of up to 4 μ s for TRY-BEN and 3 μ s for GSK-3 β . Notably, all these simulations can be executed in parallel to optimize computational efficiency.

All the work profiles obtained from SMD simulations at different pulling speeds and with different solvation models are reported below. For the TRY-BEN system, the work profiles from the 10 ns unbinding and binding simulations are represented in Figures 5.1 and 5.2, respectively. Similarly, Figures 5.3 and 5.4 display the corresponding work profiles for the 100 ns simulations, providing insight into the impact of different water FFs and longer simulation times on the binding and unbinding processes. Moreover, for the GSK-3 β system modeled with different water models, the work profiles from the 10 ns SMD unbinding and binding simulations are illustrated in Figures 5.5 and 5.6, respectively. Additionally, Figures 5.7 and 5.8 present the corresponding work profiles for the 100 ns simulations.

Then, by using our semi-automatic pipeline, we reconstruct the free energy profiles along $S(\mathbf{x})$, where $S(\mathbf{x}) = 0$ corresponds to the bound state and $S(\mathbf{x}) = 1$ represents the unbound state. The resulting profiles obtained via CFT are shown in Figure 5.9 for TRY-BEN and Figure 5.10 for GSK-3 β . The binding free energy is estimated by computing the ratio of partition functions between the bound and unbound regions, with the discriminating configuration identified as the 8th configuration along the path for TRY-BEN and the 35th configuration for GSK-3 β . Finally, the standard volume correction term is used to refine the estimates, providing standard binding

free energies that are directly comparable to experimental binding affinities.

For a comprehensive analysis, we compare the results obtained using the bidirectional CFT estimator with those from the unidirectional JE estimator. The standard binding free energies calculated using the JE for both binding and unbinding simulations across different water models are reported in Table 5.1 for TRY-BEN and Table 5.2 for GSK-3 β . According to these outcomes, the TIP3P model leads to $\Delta\hat{F}_{J,u}$ and $\Delta\hat{F}_{J,b}$ values showing the best agreement with the experimental affinity for both systems.

In the end, the effectiveness of the method is evaluated in terms of both convergence (variance) and accuracy (bias). Convergence is assessed by examining the stability of the estimates as simulation time increases, while accuracy is determined based on how well the estimates align with the experimental binding free energy upon reaching convergence.

MODEL	10 ns		50 ns		100 ns	
	$\Delta\hat{F}_{J,u}$	$\Delta\hat{F}_{J,b}$	$\Delta\hat{F}_{J,u}$	$\Delta\hat{F}_{J,b}$	$\Delta\hat{F}_{J,u}$	$\Delta\hat{F}_{J,b}$
TIP3P	-20 ± 1	-4 ± 2	-14.3 ± 0.5	-4 ± 1	-11 ± 1	-4 ± 2
TIP3P/Fw	-16 ± 1	-2 ± 1	-14.5 ± 0.8	-4 ± 2	-9 ± 1	-5 ± 1
OPC3	-28 ± 2	0 ± 3	-17.7 ± 0.9	-2 ± 2	-15 ± 1	-1 ± 1
TIP4P	-25 ± 2	-1 ± 4	-15.8 ± 1	-3 ± 3	-18.6 ± 0.9	-2 ± 2
TIP4P/Fw	-23 ± 2	0.1 ± 0.9	-13 ± 2	-2 ± 6	-11.7 ± 0.3	-2 ± 3
OPC	-22 ± 1	0 ± 2	-16.9 ± 0.6	-1 ± 1	-10 ± 2	-2 ± 2

Table 5.1: Standard binding free energies (kcal/mol) computed with the Jarzynski estimator (from unbinding and binding simulations) for TRY-BEN solvated with different water models for different simulation times.

MODEL	10 ns		50 ns		100 ns	
	$\Delta\hat{F}_{J,u}$	$\Delta\hat{F}_{J,b}$	$\Delta\hat{F}_{J,u}$	$\Delta\hat{F}_{J,b}$	$\Delta\hat{F}_{J,u}$	$\Delta\hat{F}_{J,b}$
TIP3P	-41.3 ± 0.8	5 ± 13	-21 ± 1	-3 ± 2	-15.6 ± 0.4	-7 ± 2
TIP4P	-42 ± 2	6 ± 10	-28.8 ± 0.9	2 ± 2	-24.7 ± 0.6	-4 ± 4
OPC3	-41 ± 2	8 ± 4	-35 ± 3	2 ± 4	-34.3 ± 0.6	-2 ± 3

Table 5.2: Standard binding free energies (kcal/mol) computed with the Jarzynski estimator (from unbinding and binding simulations) for GSK-3 β solvated with different water models for different simulation times.

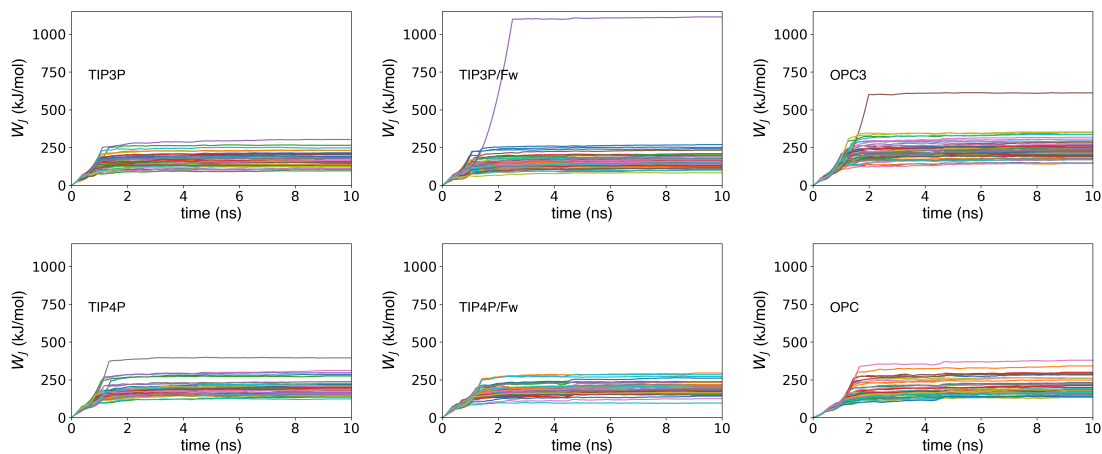


Figure 5.1: Unbinding work curves for 10 ns simulations of the TRY-BEN system solvated with different water models.

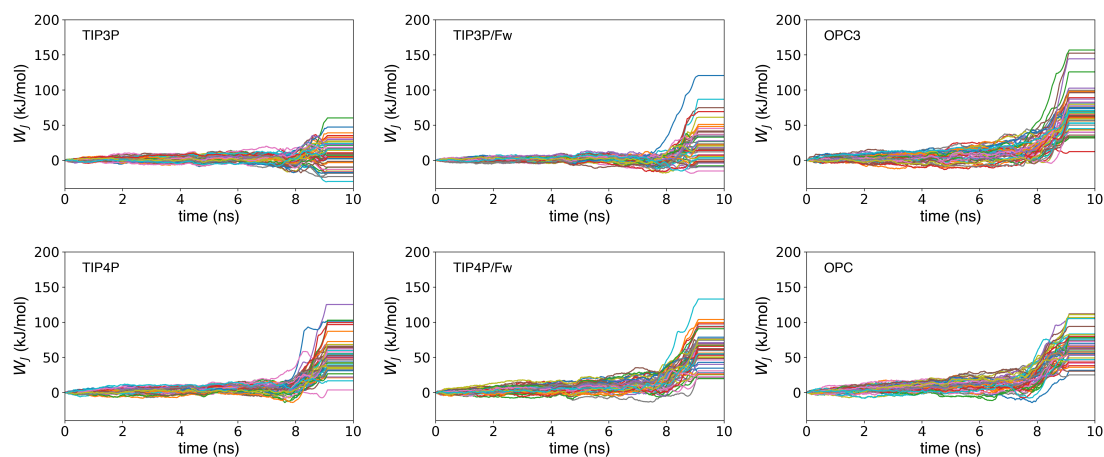


Figure 5.2: Binding work curves for 10 ns simulations of the TRY-BEN system solvated with different water models.

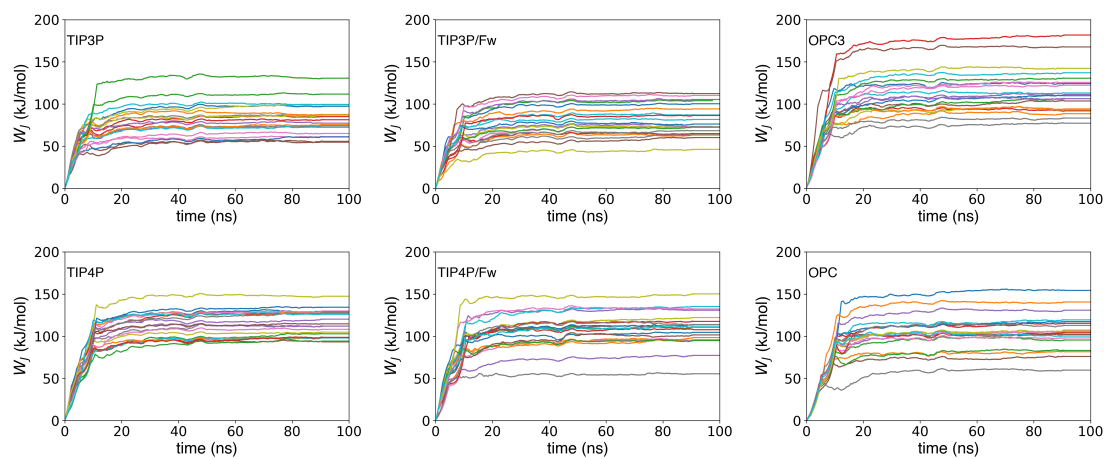


Figure 5.3: Unbinding work curves for 100 ns simulations of the TRY-BEN system solvated with different water models.

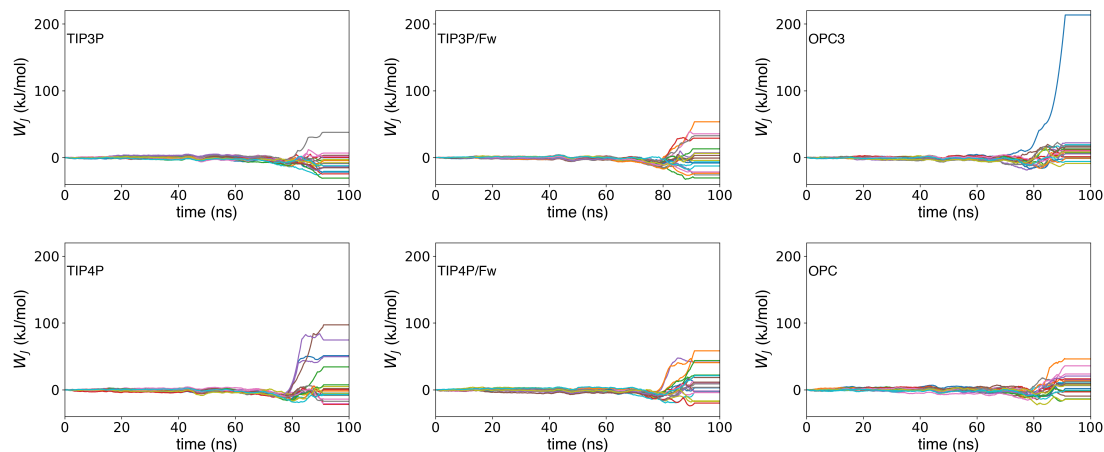


Figure 5.4: Binding work curves for 100 ns simulations of the TRY-BEN system solvated with different water models.

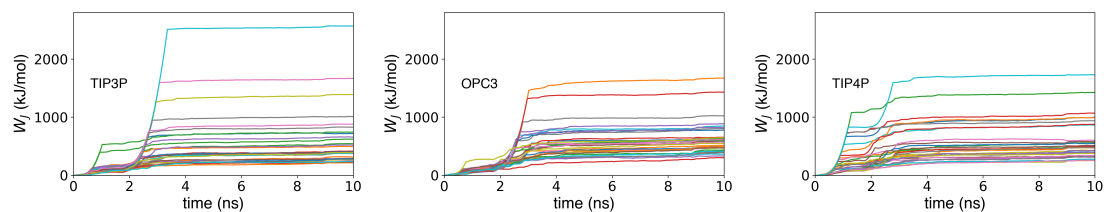


Figure 5.5: Unbinding work curves for 10 ns simulations of the GSK-3 β system solvated with different water models.

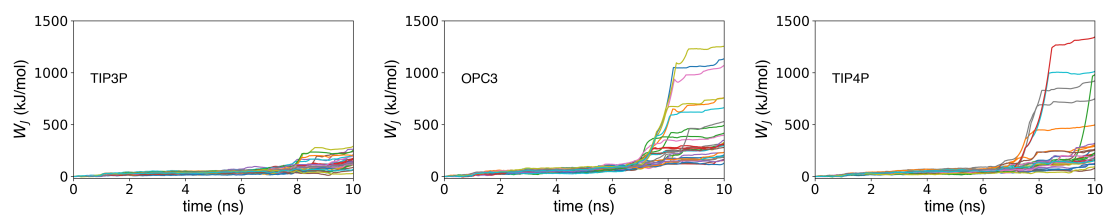


Figure 5.6: Binding work curves for 10 ns simulations of the GSK-3 β system solvated with different water models.

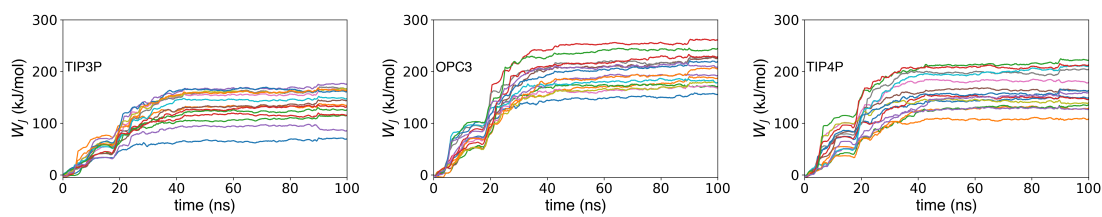


Figure 5.7: Unbinding work curves for 100 ns simulations of the GSK-3 β system solvated with different water models.

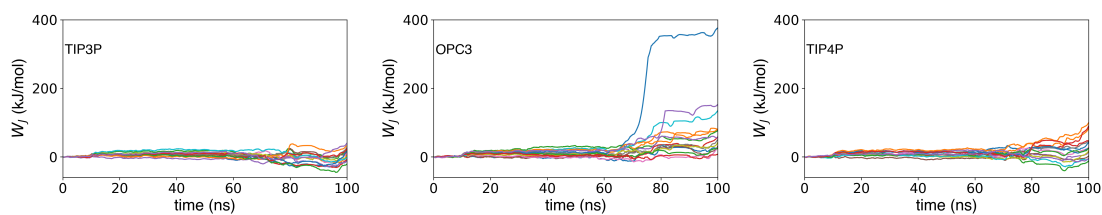


Figure 5.8: Binding work curves for 100 ns simulations of the GSK-3 β system solvated with different water models.

Impact of water models on TRY-BEN

The unbinding work profiles of TRY-BEN presented in Figures 5.1 and 5.3 show that cumulative W_J increases over time until reaching a plateau at the unbound state. Across all water models, cumulative W_J decreases as simulation time increases, as lower steering velocities reduce dissipated work. During binding simulations, the binding work profiles presented in Figures 5.2 and 5.4 show that W_J values initially remain near zero as the ligand is solvated and becomes increasingly negative over time. However, for all water FFs, a significant decrease in dissipated work is observed only in sufficiently long SMD simulations.

The influence of different water models on work profiles diminishes with longer simulations. In 100 ns SMD simulations, the dissipated work is more consistent across models. Whereas, in 10 ns simulations kinetic effects significantly impact the results, amplifying differences among water models. This is particularly evident in fast binding simulations (Figure 5.2), where the ligand must displace water molecules from the pocket, making the process highly dependent on water behavior. In contrast, unbinding is largely influenced by protein interactions, where work is mainly generated by disrupting ligand-protein contacts.

Similar trends are observed in the PMFs obtained using the CFT estimator. The choice of water model affects local heat production, influencing dissipated work and, consequently, the PMF. In fast simulations (10 ns), PMFs and binding free energy estimates vary significantly with the water model as shown in Figure 5.9 and Table 5.3. Here, kinetics has a more prominent role amplifying differences among water FFs. In contrast, longer simulations lead to more consistent results across different water models as highlighted in Table 5.3. From 100 ns simulations, CFT estimates converge to comparable standard binding free energy values.

Nevertheless, TIP4P deviates from this pattern, showing low variance in estimates regardless of pulling speed but a significant bias. Indeed, free energy estimations from all other water models show convergence towards the experimental value as the total simulation time increases, while TIP4P results seem rather invariant to different pulling speeds, demonstrating low variance across the simulation lengths.

MODEL	10 ns	50 ns	100 ns
TIP3P	-12 ± 1	-8.1 ± 0.5	-6.4 ± 0.8
TIP3P/Fw	-6.9 ± 0.9	-8.6 ± 0.7	-5.1 ± 0.8
OPC3	-11.6 ± 0.9	-7.5 ± 0.5	-6.5 ± 0.8
TIP4P	-9.7 ± 0.8	-9 ± 1	-9.2 ± 0.6
TIP4P/Fw	-8.4 ± 0.7	-7 ± 1	-5.9 ± 0.7
OPC	-7.5 ± 0.8	-7.3 ± 0.4	-4 ± 1
Experimental	-6.2		

Table 5.3: Standard binding free energies (kcal/mol) for TRY-BEN estimated using CFT for increasing simulation times compared to the experimental affinity.

Even with 10 ns simulations, TIP4P CFT-derived estimates converge rapidly toward a stable value, whereas the 3-point models require extended 100 ns simulations to reach convergence. However, the TIP4P model exhibits a noticeable bias in the estimates, deviating significantly from both the experimental values and the converged estimates obtained with other water models.

On the other hand, TIP3P and OPC3 models require longer simulation time to reach convergence. They give estimates with minimal bias, aligning closely with experimental values. Conversely, OPC, despite being one of the higher-quality water model [104], shows high variance and bias. Results from OPC simulations lead to overall unsatisfactory results in this nonequilibrium setup. Lastly, flexible models provide reasonable agreement with experimental values (no substantial bias) but come with higher computational costs, making TIP3P and OPC3 preferable options.

The findings from the TRY-BEN complex analysis leads to further investigation of the TIP3P, OPC3, and TIP4P models in a more challenging system.

Impact of water models on GSK-3 β

For the GSK-3 β complex, the work profiles obtained with different solvent models and pulling speeds exhibit trends comparable to those observed for TRY-BEN. Consequently, similar conclusions can be drawn. Analysis of the PMFs (Figure 5.10) and binding free energy estimates (Table 5.4) show that TIP3P simulations

MODEL	10 ns	50 ns	100 ns
TIP3P	-22 ± 3	-12 ± 1	-10.7 ± 0.7
OPC3	-16 ± 3	-17 ± 3	-16 ± 1
TIP4P	-19 ± 2	-13.8 ± 0.6	-14 ± 1
Experimental	-10.5		

Table 5.4: Standard binding free energies (kcal/mol) estimated using CFT for GSK-3 β increasing simulation times compared to the experimental affinity.

exhibit slower convergence and greater variance, with fluctuations in PMFs and estimates across different simulation times. Nevertheless, at 100 ns, the TIP3P estimate closely aligns with the experimental binding free energy (-10.5 kcal/mol) and shows minimal bias. In contrast, the OPC3 and TIP4P models exhibit lower variance, with their estimates stabilizing at 10 ns and 50 ns, respectively, demonstrating faster convergence. However, despite their reduced variance, their final estimates remain more biased than those obtained with TIP3P, even at 100 ns, showing a greater deviation from the experimental value. Lastly, we note that for both TRY-BEN and GSK-3 β , the binding process occurs with minimal barriers. The reason of this observation for the TRY-BEN complex may be primarily attributed to the ligand small size. Instead, for GSK-3 β , the well-optimized ABMD-derived path and the solvent-exposed nature of the binding site facilitate a smooth binding process.

Discussion

From a technical standpoint, all the examined water models are compatible with the AMBER FFs used to describe the proteins and ligands in our analysis. However, because a specific water model is usually used to develop a specific protein FF, the selection of a particular model can influence the accuracy of predictions. Furthermore, to maintain consistency and reliability in simulation results, it is crucial to ensure that also the ion parameters used for system neutralization are appropriately matched to the chosen water model.

In this study, for each water model, dissipation is evaluated from two different point of view: (*i*) in terms of cumulative dissipated work during long simulations

and (ii) in terms of dissipation variation over the total simulation length. The cumulative dissipated work during 100 ns binding and unbinding simulations is shown for TRY-BEN in Figure 5.11 and for GSK-3 β in Figure 5.12, illustrating how dissipation evolves over time. During 100 ns unbinding simulations, dissipation is relatively similar across water models, with TIP3P (both flexible and rigid) exhibiting the lowest dissipation. During binding simulations, dissipation increases towards the end as the ligand displaces water molecules to bind the target. This observation can be attributed to the significant work performed by the ligand as it displaces water molecules while approaching the protein surface. This effect is more pronounced with OPC3, OPC, and TIP4P, as these water models are more resistant to displacement by the ligand. In contrast, TIP3P/Fw and TIP4P/Fw show a more moderate increase, while TIP3P exhibits only minimal changes in dissipation.

Moreover, figures 5.13 and 5.14 illustrate the variation in dissipation with increasing simulation time for TRY-BEN and GSK-3 β , respectively. These plots evaluate the efficiency of different water models in reducing dissipation over long simulation times, providing a quantitative measure of each model's ability to mitigate energy loss during the process. Specifically, a steeper work reduction indicates a greater capacity for facilitating convergence, though rapid convergence does not always guarantee accuracy. As shown in these results, all models benefit from longer simulations, highlighting the importance of extended simulation durations. By examining the derivative of dissipated work with respect to the total simulation time, we observed that the analyzed 4-point water models are more significantly affected by longer simulations. However, although TIP3P does not show the most substantial improvement with longer simulation times, it consistently exhibits the lowest overall dissipation.

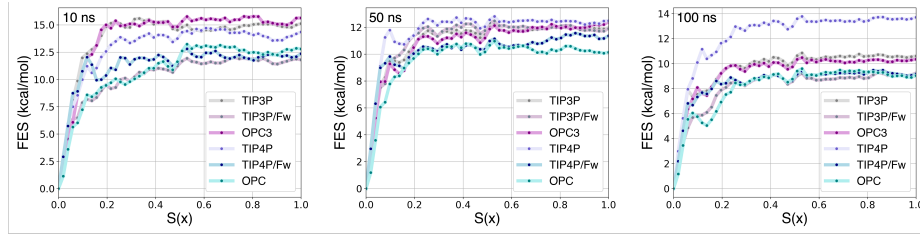


Figure 5.9: Free energy profiles along $S(x)$ ($S(x) = 0$ bound state and $S(x) = 1$ unbound state) obtained by applying CFT to varying time length SMD simulations of the TRY-BEN system solvated with different water models.

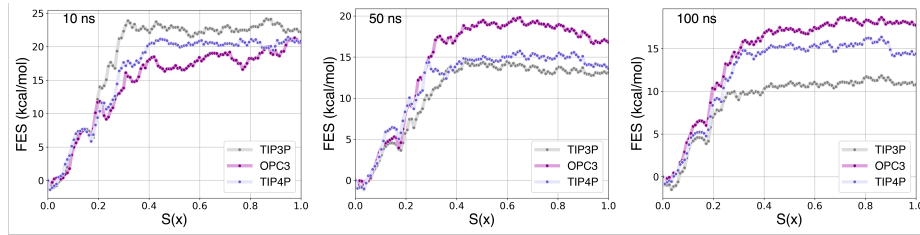


Figure 5.10: Free energy profiles along $S(x)$ ($S(x) = 0$ bound state and $S(x) = 1$ unbound state) obtained by applying CFT to varying time length SMD simulations of the GSK-3 β system solvated with different water models.

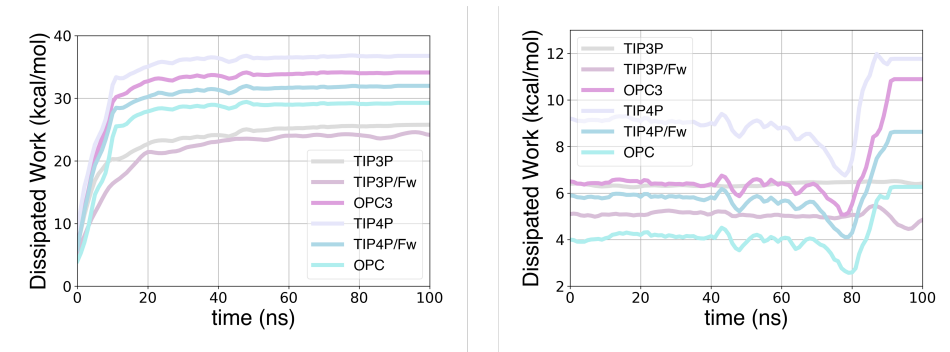


Figure 5.11: Evolution of dissipated work during 100 ns binding (left) and unbinding (right) simulations for TRY-BEN with different water models.

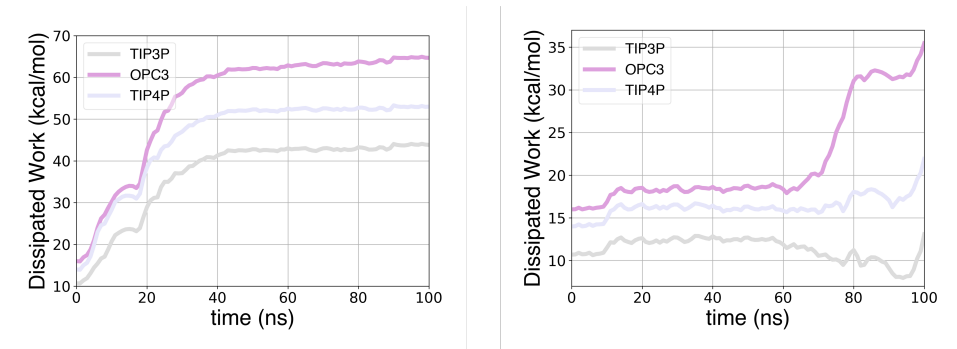


Figure 5.12: Evolution of dissipated work during 100 ns binding (left) and unbinding (right) simulations for GSK-3 β with different water models.

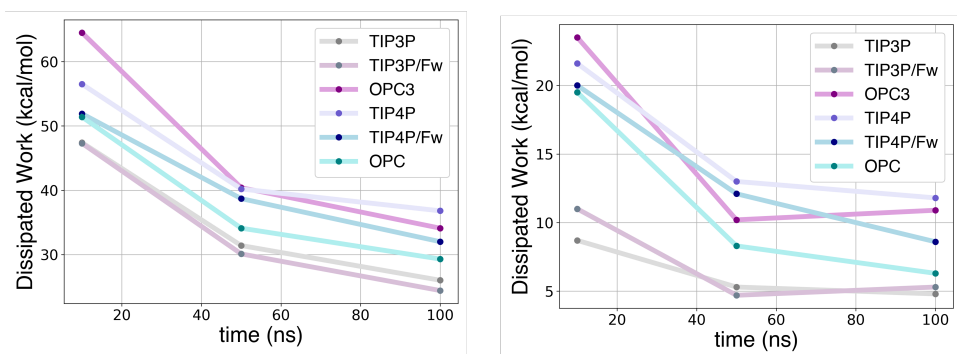


Figure 5.13: Dissipated work vs total simulation time for unbinding (left) and binding (right) simulations of TRY-BEN with various water models.

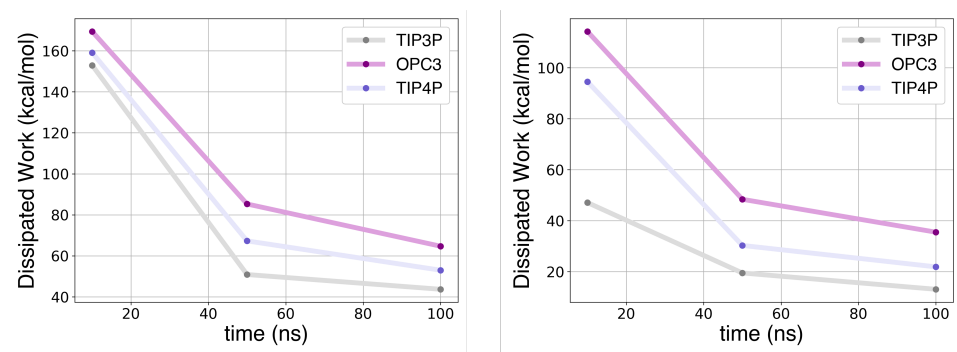


Figure 5.14: Dissipated work vs total simulation time for unbinding (left) and binding (right) simulations of GSK-3 β with various water models.

Our findings suggest that the TIP3P water model is the most effective choice for minimizing dissipation in nonequilibrium simulations while maintaining accurate

estimates for both systems studied. This advantage may be linked to TIP3P lower viscosity compared to the other water models analyzed, which facilitates water displacement and reduces dissipated work. A key reason behind this lies in its original parametrization strategy. Unlike other models, TIP3P was not parametrized to match the experimental self-diffusion coefficient, resulting in an overestimated value. In contrast, models such as TIP4P, OPC3, and OPC have diffusion coefficients closer to experimental data, leading to stronger hydrogen bonding and higher viscosity. Since TIP3P exhibits weaker hydrogen networks and lower viscosity, the ligand encounters less resistance when displacing water molecules during binding, ultimately reducing dissipation. Another factor influencing viscosity is the magnitude of partial atomic charges in water molecules. The analysed 4-point models, such as TIP4P and OPC, generally have higher absolute charge values than their 3-point counterparts, leading to increased viscosity. Among the models examined, OPC has the highest partial charges, followed by OPC3, which has the highest charges among 3-point models. Despite their accuracy in reproducing experimental viscosity, this characteristic results in greater dissipated work in nonequilibrium simulations [94].

Furthermore, as observed in the TRY-BEN and GSK-3 β complexes, when electrostatic interactions play a critical role, the dielectric constant of water becomes a key factor. Specifically, the dielectric constant influences the strength of water-water interactions, with higher values leading to weaker electrostatic forces. TIP water models were not parameterized to accurately reproduce experimental dielectric constants, resulting in an overestimation for TIP3P and an underestimation for TIP4P relative to experimental values. Consequently, TIP4P exhibits stronger electrostatic interactions, reinforcing hydrogen bonding networks and making water displacement more difficult. This explains its higher viscosity and increased dissipation compared to TIP3P.

In summary, combining the Amber protein FF with GAFF for the ligand and the TIP3P water model yields the most favorable setup for nonequilibrium simulations. TIP3P demonstrates minimal work dissipation and low estimation bias, ensuring reliable agreement with experimental data at a reasonable computational cost. De-

spite not being the most recently developed model or the best for accurately capturing hydrogen bonding networks, its lower viscosity makes it the optimal choice for practical nonequilibrium applications. Moreover, its superior performance may also stem from the fact that TIP3P was used during GAFF parametrization [215], further enhancing its compatibility in these nonequilibrium scenario.

5.5 Dissipation due to PCVs

This section aims to explore the influence of PCVs on the Jarzynski work derived from SMD simulations. In the cases of the TRY-BEN, GSK-3 β , and CB8-G8 systems, the definition of PCVs using our pipeline was relatively straightforward [P1]. This is primarily due to the moderate complexity of these systems, which involve small to medium-sized ligands binding to shallow or fully exposed pockets, with minimal degrees of freedom orthogonal to the PCVs.

However, applying the same approach to more complex systems, where substantial protein conformational changes occur, proved challenging. In such cases, defining appropriate PCVs became more difficult, leading to higher Jarzynski work values and making it harder to obtain reliable binding free energy estimates in agreement with experimental data. While extending SMD simulations could help accommodate additional degrees of freedom, this strategy would be computationally expensive for large systems. To successfully adapt our computational pipeline for real-world pharmaceutical applications, refinements are needed, particularly in optimizing the reference path and PCVs. This analysis aims to assess how PCVs impact work dissipation in nonequilibrium simulations of complex systems and to explore strategies for minimizing dissipation. This ultimately improves the convergence of binding free energy estimators.

We select the T4 Lysozyme complex with the N-phenylglycinonitrile ligand, focusing specifically on the L99A/M102Q double mutant. This less-explored variant is particularly suited for our analysis due to its deeply buried binding pocket, making it an excellent candidate for investigating how different PCVs definitions and unbinding pathways influence dissipation. Moreover, the lack of a well-defined un-

binding route also provides an opportunity to study orthogonal deviations from the reference path during SMD simulations. As seen in previously studied systems, the choice of water model can significantly affect outcomes in nonequilibrium settings. Given its lower viscosity, the TIP3P model is expected to be the most appropriate for the selected system and its buried pocket. The properties of the TIP3P model may promote more efficient water displacement during ligand binding, thereby facilitating the desolvation process within the binding pocket.

Here we present our findings for the T4 Lysozyme-N-phenylglycinonitrile complex and offer practical guidelines for the parametrization of PCVs, tailored to the specific features of the system. The goal is to enhance the convergence and reliability of nonequilibrium binding free energy calculations.

5.5.1 Reference path definition

Identifying the minimum free energy path is crucial for ensuring a low-dissipation unbinding pathway. A reference path that deviates from the real mechanism, on the other hand, can lead to high free energy barriers, resulting in significant dissipation and inaccurate binding free energy estimates. The systems studied so far have superficial or semi-superficial binding pockets, which made identifying an unbinding path relatively straightforward. However, for systems with deeply buried binding pockets, determining the most probable exit pathway for the ligand becomes more challenging [211, 12, 216]. Buried binding pockets often trap water molecules, making the exchange of water between the cavity and the bulk solvent particularly slow. This process may require significant conformational changes in both the protein and solvent network to facilitate water escape, complicating the identification of the proper pathway.

The first step of our protocol involves sampling the ligand exit using multiple 10 ns unbinding ABMD TIP3P simulations. While ABMD is a relatively mild method [135], it still requires long simulation times to accurately capture the conformational changes for buried ligands and flexible proteins, as well as the exchange of water molecules inside and outside the binding cavity. This complex scenario makes it difficult to identify a single predominant escape route. ABMD simulations help

identify meaningful pathways from a mechanistic perspective. Specifically, we focus on pathways that are associated with low energetic barriers, as these are expected to result in lower dissipation. Such paths would facilitate faster convergence when used in free energy calculations.

Differently from the previous systems, where a unique unbinding path was defined, the T4 Lysozyme ABMD trajectories reveal multiple escape routes. Here, the deeply buried pocket lacks distinct tunnels leading to the protein surface, a finding also supported by our NanoShaper investigation [201]. The NanoShaper analysis computes the Connolly-Richards molecular surface and distinguishes between exposed pockets and cryptic cavities. This pocket characterization is facilitated by our newly developed graphical interface for the original NanoShaper code, now integrated into a dedicated web server [P4]. NanoShaperWeb will be described in detail in Chapter 7. The consequence of this pocket depth is that the protein must undergo conformational "breathing" and structural fluctuations to facilitate the ligand entry or exit from the binding cavity [12]. Additionally, the two-point mutations (L99A and M102Q) make the cavity polar, adding further complexity as the ligand molecules must be properly oriented to form hydrogen bonds with Gln102.

ABMD simulations and subsequent analysis reveal two predominant unbinding paths, designated as Path 1 and Path 2. Among these, Path 2 appears to be the most probable, although Path 1 still maintains a non-zero likelihood of occurring. To account for both possibilities, we apply the path algorithms to both routes obtaining the two final reference paths depicted in Figure 5.16. Path 1 and Path 2 are composed of ligand atoms only, with 49 and 73 equidistant conformations, respectively (the C_α atoms of the protein pocket are used for alignment before calculating the RMSD). As shown by the pink-colored path in Figure 5.16, Path 1 involves the ligand escaping from the pocket by passing between the lower regions of helices D and G. This route is consistent with observations from the single L99A mutant [216]. In contrast, the light-blue path in Figure 5.16 illustrates Path 2, where the ligand exits through a region between helices D and G but closer to helix F. This pathway requires a significant conformational rearrangement

in the upper part of the buried binding site, particularly involving substantial motions of helix F (residues 108–114), which expands the cavity and facilitates ligand unbinding. It is important to note that the primary goal of the ABMD simulations is to reveal mechanistically relevant unbinding pathways. Particular attention is given to those with low energy barriers and minimal dissipation, as these features enhance the convergence of free energy calculations. While our investigation identified two main routes that meet these criteria, it is relevant to acknowledge that additional alternative binding or unbinding trajectories may exist for the T4 Lysozyme L99A/M102Q mutant, as suggested in previous studies [217, 216].

Following our nonequilibrium protocol, for the two reference paths, 25 bidirectional PCVs-SMD simulations are run, each lasting 100 ns, totaling 5 μ s of simulation time per path. We then reconstruct the Jarzynski work profiles for each simulation series and we apply the CFT estimator to reconstruct the PMFs. Subsequently, the standard binding free energies, comparable with the experimental value of -5.52 ± 0.18 kcal/mol, are obtained.

For Path 1, the work curves are reported in 5.17, and the final ΔF_b° estimate is -4 ± 1 kcal/mol. Additionally, we calculate the dissipated work for both forward and backward transformations, which is determined as the difference between the average W_J and the CFT estimate $\Delta \hat{F}_{CFT}$. The binding free energy and the corresponding work dissipation for Path 1 are reported in Table 5.5. In contrast, the work curves for Path 2 (shown in the upper section of Figure 5.18) reveal an unexpected behavior. When the system reaches a partially unbound conformation ($S(x)$ values around 39 and 40), the ligand remains attached to the protein surface and fails to transition toward the unbound state. Visual inspection of the trajectories indicates that the ligand becomes trapped due to multiple contacts with side chains of residues from helix F and neighboring regions (especially residues: Arg125, Arg119, Asn116, Met120, and Gln123). Because of these interactions, the ligand is not able to follow the moving SMD center synchronously. As a result, the Jarzynski work, which depends on the distance between the ligand and the center of the bias, increases exponentially, hindering the accurate calculation of the

binding free energy. This issue may arise because Path 2 involves significant conformational rearrangements in the binding pocket to allow the ligand to fully exit, particularly with helices F and G. In system with deeply buried binding sites, our finding (further validated through Well-Tempered MetaD analysis) underscores the importance of pocket flexibility and rearrangements. The details of Well-Tempered MetaD analysis are reported in the Supporting Information (Chapter 9).

The challenges encountered highlight the need to adjust our approach for defining the reference path for more intricate systems. Even if the ligand path appears mechanistically accurate, our original strategy (successful for simple systems) do not capture the degrees of freedom of the protein in the reference path, which however are relevant to the binding mechanism. In simpler cases, the SMD bias only influences the ligand atoms along the optimized path, while the protein may behave differently from what the actual binding mechanism requires. To address this issue, it becomes clear that relevant protein degrees of freedom must be included in the reference path definition. By incorporating both the ligand and protein pocket atoms into the path, we ensure that during SMD simulations the bias potential would not only guide the ligand toward the binding pocket but also support the necessary transitions of the protein between open and closed conformations. Therefore, using the same ABMD trajectories, we apply the refined approach to generate new reference paths that include both the ligand and pocket heavy atoms (within 6 Å of the ligand). As a result, the updated optimized paths now consist of 20 and 29 conformations for Path 1 and Path 2, respectively.

By utilizing these refined paths, we run 25 binding and 25 unbinding replicas of SMD. According to the resulting work profiles (lower section of Figure 5.18), incorporating pocket atoms into Path 2 significantly reduced the generated work. This adjustment allows us to estimate the binding free energy for both paths using the CFT estimator. These estimates, along with the dissipated work for both paths relative to their converged CFT values, are presented in Table 5.5.

The differences in dissipation and binding free energy across different path definitions highlight the sensitivity of the results to the definition of the PCVs for this complex system. While including the pocket improves the results for Path 2, it had

PATH	ΔF_{CFT}	$W_{diss,bind}$	$W_{diss,unbind}$
Path 1	-4 ± 1	32	40
Path 2	*	*	*
Path 1 with Pocket	-2 ± 4	55	29
Path 2 with Pocket	-8 ± 2	29	58
Path 1 with Pocket and smaller Z	-7.8 ± 0.9	8	32

Table 5.5: Standard binding free energies (kcal/mol) estimated using CFT and dissipated work values (kcal/mol) for 100 ns SMD simulations of the T4 Lysozyme using different paths. Values for Path 2 are missing as the high dissipated work prevented their calculation.

the opposite effect for Path 1. The new version of Path 1 (including pocket atoms) shows an increase in dissipated work during binding, which causes the binding free energy estimate to diverge further from the experimental value. The binding work profiles for the updated Path 1 (upper section of Fig. 5.19), reveal a noticeable increase in W_J upon reaching the bound state.

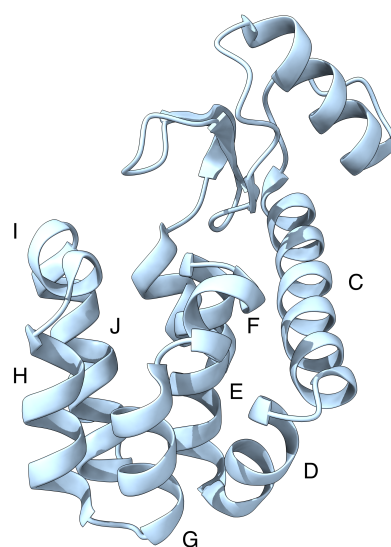


Figure 5.15: Crystal structure of the T4 Lysozyme (PDB ID: 2RBN). Helices involved in ligand binding are labeled from C to H, following Ref. [217].

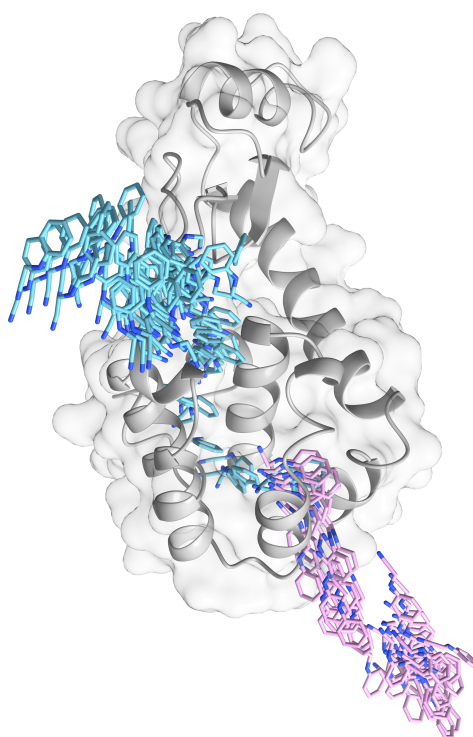


Figure 5.16: Reference paths for T4 Lysozyme in complex with N-phenylglycinonitrile: Path 1 in pink and Path 2 in light blue.

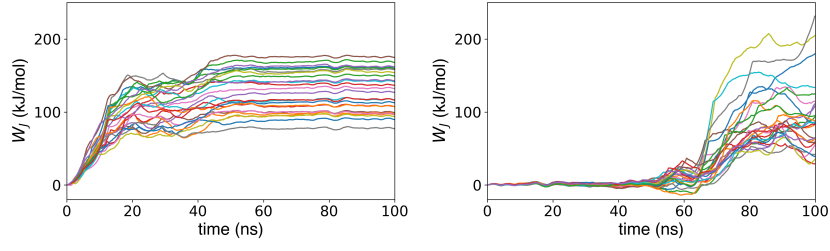


Figure 5.17: Unbinding (left) and binding (right) work curves of 100 ns for the T4 Lysozyme system with Path 1 defined by using ligand atoms only.

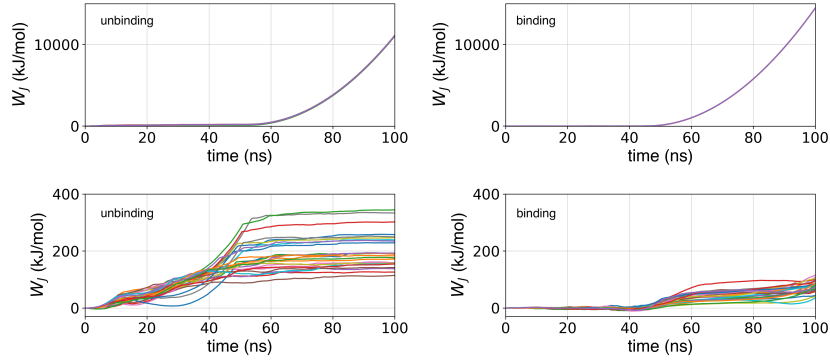


Figure 5.18: Jarzynski work profiles for unbinding and binding simulation of 2RBN using Path 2 with only ligand atoms (upper section) and refined Path 2 including ligand and pocket atoms (lower section).

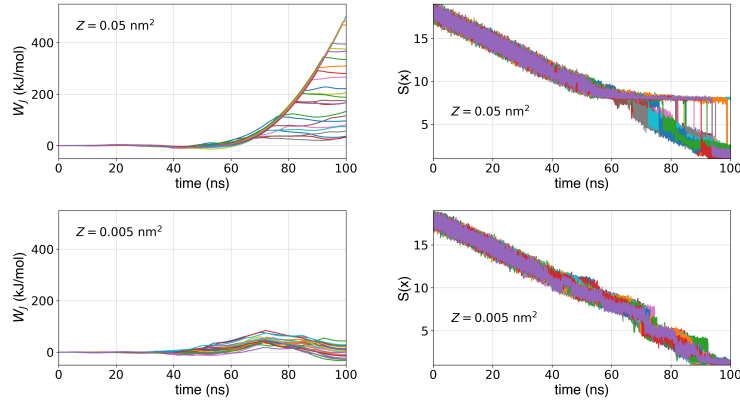


Figure 5.19: Jarzynski work profiles and corresponding evolution of $S(x)$ for binding simulation of 2RBN using Path 1 with ligand and pocket atoms with an upper wall centered on $Z = 0.05 \text{ nm}^2$ (upper section) and $Z = 0.005 \text{ nm}^2$ (lower section).

5.5.2 Orthogonal deviation from the path

Considering Path 1, the results remain suboptimal despite efforts to refine the definition of the PCVs. Upon visual inspection, we observe that in some binding trajectories, the PCV $S(x)$ struggles to closely follow the center of the SMD bias $\hat{S}(t)$, which evolves linearly with time, as shown in the upper section of Figure 5.19. As a result, in these trajectories the stiff-spring approximation is not satisfied, meaning that the evolution of $S(x)$ is not in synchrony with that of $\hat{S}(t)$. The reduced spring stiffness k could be one possible cause of this observation. While increasing this parameter may help the system better track $\hat{S}(t)$ [218], doing so can lead to excessively high values of W_J introducing undesirable dissipation. To resolve this issue, it is essential to ensure that the evolution of the system aligns with the reference path (following the SMD center) avoiding getting trapped on the protein surface. For this reason, our approach focuses on controlling the orthogonal deviation of the system from the path, specifically targeting the PCV Z . Thus, during the SMD simulations, we apply an upper wall along Z to prevent significant deviations from the reference path. Previously, the threshold for this upper wall was set at $Z = 0.05 \text{ nm}^2$, meaning the harmonic restraint becomes active only when the orthogonal deviation Z exceeds 0.05 nm^2 . This threshold was chosen to allow for proper trajectory progression while still accommodating ligand conformational changes. However, in this case, this threshold of the harmonic wall appears too loose. $Z = 0.05 \text{ nm}^2$ permits the ligand to deviate too far from the reference path. To address this, we recommend lowering the threshold at which the wall on Z becomes active to maintain tighter control over the system evolution.

Bidirectional SMD simulations are repeated using the same Path 1 (without modifying S) but with a decreased threshold for the upper wall, hence $Z = 0.005 \text{ nm}^2$. As illustrated in Figure 5.19 (lower section), this adjustment significantly lowered the work required during binding simulations. Figure 5.19 also depicts the $S(x)$ evolution over time, confirming the correct behavior of $S(x)$. Now, $S(x)$ evolves linearly over time, closely following $\hat{S}(t)$, underling better synchronization between the system evolution and the reference path (i.e., $S(x)$ closely follows $\hat{S}(t)$). Using these optimized simulations, we reconstruct the PMF and estimate the binding free

Path	JE binding	JE unbinding	CFT
Path 1	-1 ± 10	-22 ± 0.8	-10 ± 3
Path 2	*	*	*
Path 1 with Pocket	1 ± 12	-15 ± 1	-2 ± 4
Path 2 with Pocket	2 ± 7	-20 ± 4	-8 ± 2
Path 1 with Pocket and smaller Z	-4 ± 1	-10 ± 1	-7.8 ± 0.9

Table 5.6: Standard binding free energies (kcal/mol) estimated using the Jarzynski and CFT estimators for 100 ns SMD simulations of the T4 Lysozyme using different paths. Values for Path 2 are missing as the high dissipated work prevented their calculation.

energy, yielding $\Delta F_b^\circ = -7.8 \pm 0.9$ kcal/mol. Notably, the dissipated work during binding simulations with $Z = 0.005$ nm² is substantially lower than the dissipated work with $Z = 0.05$ nm², as detailed in Table 5.5. The binding free energy estimate obtained using Path 1 with $Z = 0.005$ nm² aligns more closely with the estimate from Path 2, which incorporates the pocket degrees of freedom. In contrast, all trajectories generated with Path 2 (including the pocket) are already within the stiff-spring approximation, eliminating the need to tighten the upper wall further. For a complete comparison, the results of all T4 Lysozyme paths obtained using the Jarzynski estimator are reported in Table 5.6. Results of Path 2 are not reported because the high dissipated work prevented their calculation.

Overall, our findings emphasize the critical role of properly defining PCVs in binding free energy calculations, particularly when applied to complex systems such as T4 lysozyme. All relevant degrees of freedom (as protein rearrangements) must be considered when constructing reference paths. Additionally, adjusting the activation threshold of the harmonic wall for the PCV Z based on system size and behavior is essential to ensure the system remains within the stiff-spring approximation. By including these refinements, T4 Lysozyme results become better, albeit not in perfect agreement with the experimental affinity. In the literature, it is largely recognized that systems with buried pockets (such as T4 Lysozyme) pose a significant challenge when using geometric CV approaches, with alchemical routes being more suitable for buried cavities [211]. Compared to alchemical methods, path-based methods are in general more computationally demanding and require higher human intervention. However, systems with deeply buried pockets

can be challenging also for alchemical methods. In fact, buried protein pockets may require very long simulation times to be completely solvated after the ligand annihilation. Additionally, path-based methods provide more interesting mechanistic information than alchemical ones. They can be the preferred choice in highly intricate systems, for instance in protein-protein binding free energy calculations [112]. This empathize the actual importance of geometrical routes in biophysical computations [112].

Our refinements push the practical limits of the method, demonstrating that, even for highly complex systems, it is possible to develop protocols with reduced dissipation while maintaining mechanistic accuracy.

5.6 Conclusion

The two primary challenges faced in nonequilibrium simulations are: exhaustively sampling the transformation process and accurately computing the free energy difference from these samples. While CFT efficiently addresses the latter, challenge of optimal sampling remains unresolved. In this context, minimizing dissipated work is crucial, as identifying the most efficient transformation paths leads to minimally dissipative routes.

Here, we present practical guidelines for refining our semi-automated protocol for free energy calculations, with an emphasis on minimizing dissipation. Our results highlight the significant impact of both the water model and the parametrization of PCVs on the convergence rate of nonequilibrium estimators. Notably, despite its lower accuracy in representing pure water, the TIP3P model exhibited superior convergence when paired with AMBER protein FFs in two-sided nonequilibrium simulations. This can be attributed to TIP3P lower viscosity, which reduces heat transfer during simulations and decreases dissipation, thereby speeding up convergence in free energy calculations. These conclusions are validated using two model systems, namely TRY-BEN and GSK-3 β .

In the third test complex (T4 Lysozyme), we do not achieve chemical accuracy. However, we successfully optimized the reference path and PCV definitions to

minimize dissipation in nonequilibrium strategies. Including protein pocket degrees of freedom in the reference path proved to be highly beneficial. Our results highlight the difficulties posed by buried binding pockets, as even MetaD struggled to provide accurate free energy estimates. Consistent with literature, systems with buried pockets present substantial challenges when using geometric CV approaches. Nonetheless, through careful refinement of PCV definitions, we maximized the practical utility of the method.

We hope that the presented guidelines will support the broader application of nonequilibrium path-based approaches in drug design, as these methods offer significant promise for future research and practical applications in the field.

Chapter 6

Expanding the applicability of nonequilibrium binding free energy estimation: from protein to RNA [P3]

6.1 Aim and Objectives

Following the definition of practical guidelines to enable broader application of nonequilibrium techniques in drug discovery (Chapter 5); in this Chapter we aim to apply our refined nonequilibrium strategy to tackle the challenge of accurately estimating binding free energies in biologically and pharmaceutically relevant systems, including both protein-ligand and RNA-ligand complexes. As representative case studies, we investigate the binding of the anticancer drug Gleevec to Abl-tyrosine kinase, as well as the interaction of two distinct ligands with the preQ1 RNA riboswitch. The Abl-Gleevec system presents challenges related to its large molecular size and the significant conformational transitions of the Abl-tyrosine kinase, whereas the RNA-based systems demand strategies capable of capturing substantial structural flexibility. To overcome the specific complexities of these systems, we further refine our methodology based on nonequilibrium SMD and

PCVs. Overall, this work aims to expand the applicability and demonstrate the robustness of our path-based SMD approach in handling structurally dynamic and complex biomolecular targets, offering a reliable framework for predicting binding free energies across diverse classes of molecular systems.

6.2 Introduction

Building on our previous work on ΔF_b estimation using PCVs combined with nonequilibrium approaches (successfully applied to systems of moderate complexity) we now turn our attention to a key unresolved challenge: the Abl-Gleevec complex. This system resulted particularly complex due to the high amount of work dissipation observed during simulations, which leads to poor convergence in free energy estimates (Section 4.4.3 and Reference [P1]). The difficulties associated with accurately estimating binding free energies stems from multiple factors, including the high number of degrees of freedom and the substantial conformational flexibility of both the receptor and the ligand. To enhance convergence in large and flexible systems, we previously developed a refined strategy specifically designed to reduce dissipative effects during nonequilibrium simulations (Section 5 and Reference [P2]).

In the present work, we apply this enhanced nonequilibrium methodology for binding free energy estimation to intricate biomolecular systems, including both protein and RNA targets. We examine the Abl-tyrosine kinase bound to Gleevec and two RNA-ligand complexes, demonstrating the flexibility and robustness of our computational framework. For the RNA systems in particular, we introduce a tailored strategy that leverages electrostatic interactions to construct reliable unbinding pathways.

6.2.1 Systems

The systems selected for this study exhibit distinct and noteworthy features that need detailed discussion. Here, we provide an analysis of these two biologically and pharmaceutically relevant complexes, highlighting their complementary challenges.

On one hand, the Abl-Gleevec complex poses difficulties primarily due to significant conformational rearrangements of the protein, driven by the size and complexity of the ligand. On the other hand, the RNA-ligand systems are challenging mainly because of the inherent flexibility of the RNA receptor.

Abl-Gleevec

Imatinib, a 2-phenylaminopyrimidine derivative marketed under the name Gleevec, is an anticancer drug designed to mimic the structure of ATP, the natural substrate of Abl kinase. This structural similarity enables it to occupy the ATP-binding pocket of the kinase, thereby inhibiting ATP access and effectively blocking the enzymatic activity of the protein.

From a computational standpoint, accurately modeling the binding and unbinding processes of Imatinib to Abl kinase presents substantial challenges. These arise primarily from the extensive conformational changes of Abl kinase required for the transition between the active state (predominant in solution) and the inactive conformation stabilized upon Imatinib binding [219]. To illustrate these extensive conformational changes, Figure 6.1 presents the X-ray crystallographic structures of c-Abl kinase in the two distinct states. The top panel shows the active conformation (left), characterized by an open activation loop (A-loop) and the DFG motif in the “in” position (PDB ID: 2F4J), and the inhibited conformation (right), where the A-loop is closed and the DFG motif adopts the “out” conformation upon binding to Gleevec (PDB ID: 1IEP). One of the most critical structural transitions in Abl kinase involves the A-loop, which regulates access to the catalytic site and includes a tyrosine residue whose phosphorylation modulates enzymatic activity [220]. In the active form of Abl, the A-loop adopts an open conformation that facilitates substrate (ATP or ADP) binding and subsequent phosphorylation. Upon Imatinib binding, however, the A-loop undergoes a substantial structural rearrangement, including a rotation that displaces central residues by as much as 35 Å. Another important structural element is the conserved DFG (Asp-Phe-Gly) motif within the A-loop. In the active, uninhibited conformation, the DFG motif adopts the “DFG-in” orientation, with the aspartate residue positioned to coordinate with

Mg-ATP in the active site. In contrast, in the inhibited conformation stabilized by Imatinib, the DFG motif rotates approximately 180° , shifting to the “DFG-out” position. This flip repositions the Phe side chain into the catalytic pocket, effectively blocking substrate binding and enzymatic activity.

The therapeutic effectiveness of Imatinib is closely related to the ability of Imatinib to induce extensive protein conformational changes, affecting over 20 residues. Moreover, the considerable flexibility and substantial size of Imatinib further complicate computational modeling of the binding and unbinding processes. Despite these challenges, studying the interaction dynamics of the Abl-Gleevec complex remains highly relevant due to its biological and pharmaceutical importance [221]. To address this, we apply our nonequilibrium protocol, which leverages physically relevant unbinding pathways, as a suitable framework for estimating the binding free energy of this complex. This approach not only facilitates accurate free energy calculations but also yields mechanistic insights into the binding process. As such, the Abl-Gleevec complex serves in this work as a prototypical example of a pharmaceutically important, yet computationally demanding, protein-ligand system.

Riboswitch-ligand

The complexity observed in the Abl-Gleevec system is similarly reflected in the challenges of targeting RNA with small molecules. RNA has recently emerged as a highly attractive pharmaceutical target, driven by its critical biological roles and structural versatility [222, 223, 224, 225, 226, 227]. This interest has been further reinforced by the FDA approval of Risdiplam, the first small-molecule drug designed to modulate RNA function [228, 229].

RNA molecules present unique challenges as drug targets due to their complex structural dynamics, high charge density, and the limited understanding of RNA-ligand interactions [230, 231, 232]. As a result, further exploration into RNA-ligand recognition is crucial, as it could drive significant progress in the development of small-molecule drugs targeting RNA. In particular, RNA riboswitches [233] have gained attention as promising drug targets [234, 235, 236], due to their ability to bind metabolites and regulate gene expression. Their importance is amplified by

their prevalence in bacteria, including human pathogens, making them especially relevant in the context of the growing threat of bacterial resistance, as highlighted by the World Health Organization (WHO) [234, 235]. A notable example is the ongoing efforts to target the bacterial FMN riboswitch, with a compound currently in preclinical stages [237]. With available experimental structures of RNA-riboswitch-ligand complexes and corresponding affinity data [238, 239], we use the preQ1 riboswitch as a test case to demonstrate our method for RNA-ligand binding studies.

6.3 Methods

The key steps of our nonequilibrium strategy, which integrates SMD, PCVs, and the CFT, are outlined in Section 4 and summarized in Figure 4.2. As anticipated, the approach utilizes multiple parallel SMD replicates to sample the binding and unbinding processes along a predefined physical pathway that connects the bound and unbound states. The definition of this reference path is especially critical for the highly complex system examined here, warranting particular emphasis on the initial stages of the pipeline. Then, by performing bidirectional simulations, the CFT enables the calculation of the PMF using data from both binding and unbinding trajectories.

6.3.1 ABMD and PCVs

The first step of the workflow involves defining the PCVs by running multiple ABMD trajectories of ligand unbinding from the native bound pose, as resolved in the corresponding X-ray crystallographic structures. For protein-ligand systems, unbinding is initiated from the bound complex and sampled using ABMD, where the bias is applied along the defined electrostatic-like CV. An ABMD force constant of $10^{-17} \text{ (kJ/mol)}^{-3}$ is employed in all simulations.

In contrast, for the RNA-ligand complexes, we leverage the intrinsic electrostatic properties of the systems. Due to the strongly negative electrostatic potential of RNA and the positively charged nature of the ligands, ABMD simulations are per-

formed using a CV based on the Debye-Hückel interaction energy. This approach effectively steers the system toward a state of negligible electrostatic interaction between the RNA and the ligand, resulting in complete dissociation. Force constants of 10^{-1} and 10^{-2} (kJ/mol) $^{-3}$ are employed during these ABMD simulations.

Multiple ABMD simulations are run for each complex to explore the diversity of all possible unbinding pathways. Among these, the trajectory occurring with the highest frequency is chosen as the initial guess for the unbinding path definition. The selected trajectory is then refined using the two path optimization algorithms (i.e., the Principal Path algorithm and an enhanced version of the Equidistant Waypoint algorithm). The resulting equidistant, optimized path is then employed to perform bidirectional SMD using PCVs. PCVs enable an unambiguous mapping of each configuration position relative to the predefined unbinding path within the configurational space.

6.3.2 SMD and PMF estimation

To sample the binding and unbinding events, two-sided SMD simulations with PCVs are run. Multiple nonequilibrium trajectories are generated at a defined pulling speed, which governs the motion of the SMD restraint along the path. As the restraint moves at a constant velocity, the duration of each simulation is inversely related to the pulling speed. The pulling speed and the number of SMD replicas are selected to ensure the convergence of free energy estimates. For all complexes studied, 30 binding and 30 unbinding SMD trajectories of 100 ns each are required. To assess robustness across different pulling regimes, additional SMD simulations of the Abl-Gleevec complex are conducted with extended durations of 200 ns. A moving harmonic restraint with a force constant of 20 kJ/mol is applied in all cases. Additionally, to maintain the system along the path, a half-harmonic (flat-bottom) potential is also applied along Z PCV. This upper wall activates only when the $Z(x)$ value exceeds a threshold of 0.05 nm², preventing excessive deviation from the reference path.

During SMD simulations, the primary quantity of interest is the Jarzynski work (Eq. 2.60), which reflects the work done on the system along the nonequilibrium

path. Notably, increasing the pulling speed leads to a higher total W_J , which in turn results in greater dissipated work (Eq. 2.61), thereby affecting the accuracy and convergence of free energy estimates.

Subsequently, the free energy profiles along $S(x)$ are reconstructed from the SMD simulations using the bidirectional nonequilibrium CFT estimator, as defined in Equation 2.63. By incrementally solving Equation 2.64 for increasing values of i , the complete free energy profiles along $S(x)$ are obtained. Here, $S(x)$ is normalized between 0 and 1, with $S(x) = 0$ representing the bound state and $S(x) = 1$ corresponding to the unbound state. Finally, for comparison with the bidirectional estimates, unidirectional free energy calculations are also performed using the JE. This approach involves applying the same nonequilibrium protocol to unidirectional trajectories, where the system evolves from the initial to the final state (identical protocol). The free energy is then estimated by averaging the exponential of the negative work, following the Jarzynski relation (Eq. 2.62).

6.3.3 Standard Binding Free Energy Estimation

To enable comparison between our computational results and experimental data, we calculate the standard binding free energy (ΔF_b°), which includes both the computed binding free energy (ΔF_b) and a standard-state volume correction term (ΔF_v), as outlined in Equation 4.4. This approach involves selecting the value of the PCV $S(x)$ that delineates the transition between bound and unbound states based on analysis of the PMF and visual inspection of the trajectories. Finally, uncertainties in the binding free energy estimates are assessed using bootstrap resampling analysis, as outlined in our previous methodology [P1].

6.3.4 Well-Tempered MetaDynamics simulations

To benchmark our results against established methodologies [240, 241, 242, 179], we run WT-MetaD simulations with PCVs. During these enhanced simulations, Gaussians are deposited every 250 MD steps, using a bias factor of 15. The initial Gaussian height is set to 1.0 kcal/mol, with widths of 0.2 along $S(x)$ and 0.01 nm² along $Z(x)$. Consistent with the SMD protocol, a restraining wall is applied along

the orthogonal coordinate Z at a threshold of 0.05 nm^2 , using a force constant of $4 \cdot 10^7 \text{ kJ mol}^{-1} \text{ nm}^{-4}$. The production runs yielded $1 \mu\text{s}$ of simulation for each riboswitch-ligand system and $1.5 \mu\text{s}$ for the Abl-Gleevec complex. Convergence is assessed using two main criteria: (i) complete exploration and diffusivity along the $S(x)$ collective variable and (ii) reduction of the deposited Gaussian height to below 10% of its initial value, consistent with established practices in the laboratory [180].

For each system, the free energy profile from WT-MetaD is reconstructed, and the global minimum (corresponding to the bound ligand state) is used as the reference point and set to zero. The standard binding free energies are then computed from the free energy profile using the same approach employed for the SMD-based estimates. Finally, to quantify statistical uncertainty, bootstrap analysis is performed by dividing each simulation into 10 blocks and conducting 400 resampling iterations.

6.3.5 Setup of the systems

Abl-Gleevec

The Abl kinase structure features a compact N-terminal lobe, made up of five β -strands and one α -helix, and a more expansive C-terminal lobe that contains several α -helices. These two domains form a cleft that serves as the binding pocket for ATP. Figure 6.1 displays the crystallographic structure of the Abl-Gleevec complex and contrasts it with the active form of Abl kinase, highlighting the structural differences between the ligand-bound and active conformations.

The binding selectivity and mechanism of Abl kinase are mainly influenced by two critical structural elements: the A-loop (residues 384-403) and the glycine-rich loop (P-loop, residues 248-257). The A-loop, beginning with the conserved DFG motif, is essential for modulating the kinase conformational state, adopting a folded structure in the inactive form and extending outward when the kinase is active. Meanwhile, the P-loop (also known as the phosphate-binding loop), located in the N-terminal lobe, facilitates ATP binding mainly through hydrophobic interactions.

Additionally, the gatekeeper residue Thr338 plays a key role in controlling access to the active site, further contributing to ligand selectivity and kinase regulation.

The Abl-Gleevec complex is modeled using the X-ray crystallographic structure available under PDB code 1IEP (chain A). In this conformation, the A-loop within the C-terminal lobe adopts a closed conformation, and the conserved DFG motif is observed in the DFG-out arrangement, characteristic of the inactive kinase state. Gleevec binds within the cleft between the amino- and carboxy-terminal lobes, stabilizing this inhibitor-bound configuration. Key hydrophobic interactions are formed with two conserved residues, Leu248 and Val256, which contribute significantly to ligand binding. The amide group of Gleevec plays a crucial anchoring role by forming hydrogen bonds with nearby Glu and Asp residues, guiding the ligand into its optimal orientation. Moreover, the secondary amine's NH group engages in a specific interaction with the side chain of the gatekeeper residue, further reinforcing binding stability. The AMBER99SB FF [71] is used to model Abl kinase, while the GAFF and the AM1-BCC point charges are used for Gleevec. Assuming standard pK_a values, the protonation states of amino acids are assigned according to their expected ionization states at physiological pH. Moreover, the piperazine moiety of Gleevec is known to carry a positive charge when bound within the Abl active site, enhancing its interactions with nearby residues. Therefore, the outermost N-methyl nitrogen of the piperazine ring is modeled in its protonated form, consistent with previous computational and structural studies [243, 221]. Based on our previous findings [P2], the system is solvated using the TIP3P water model, with the complex placed in a cubic box measuring 1.5 nm per side to ensure sufficient space for complete ligand dissociation from the protein binding site. To mimic physiological conditions, sodium and chloride ions are added to achieve a salt concentration of 0.15 M, ensuring system neutrality. The solvated protein-ligand complex then undergoes energy minimization and equilibration, following our standard protocol [P1].

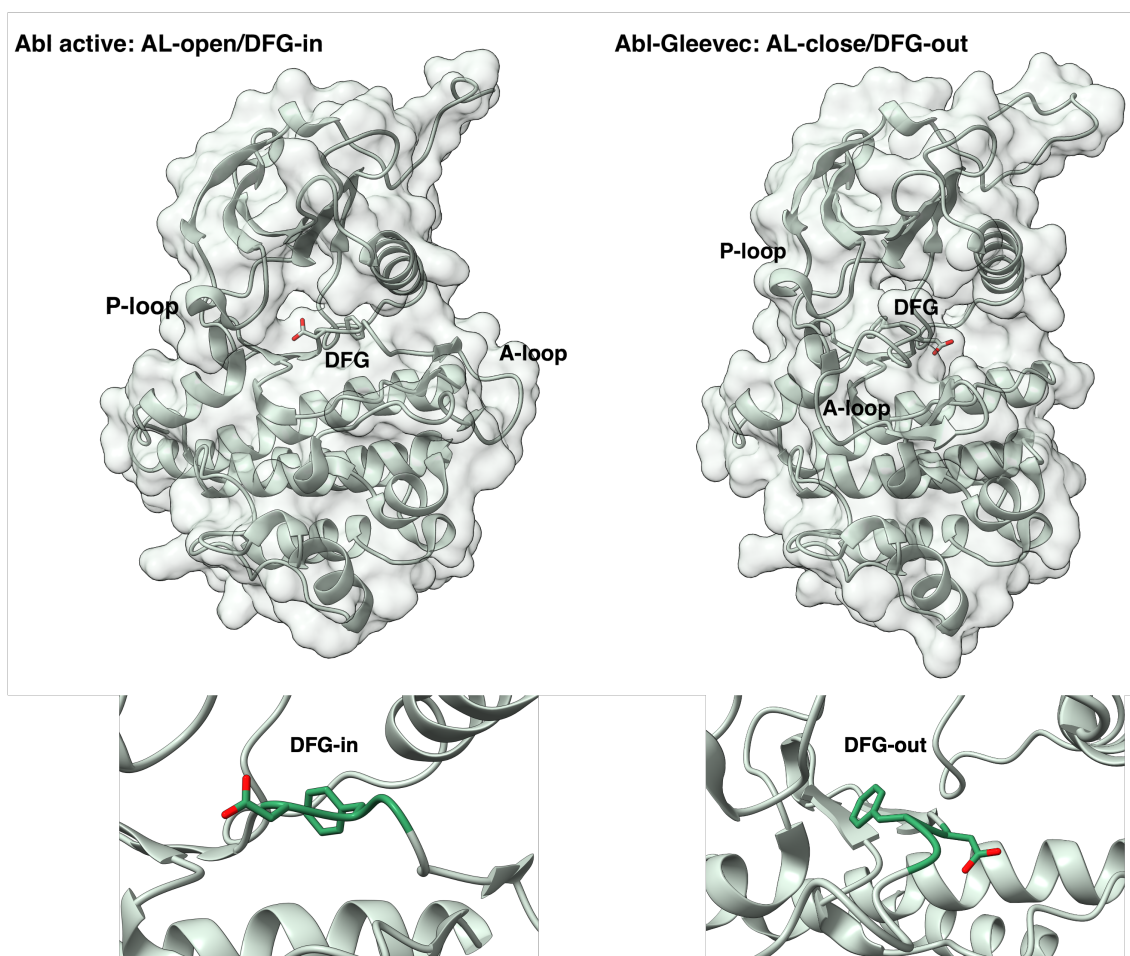


Figure 6.1: (Top) X-ray crystallographic structures of the c-Abl kinase (left) in an active state (A-loop in the open conformation and DFG-flip in the "in" conformation, PDB ID: 2F4J) and (right) in an inhibited state, in complex with the ligand Gleevec (A-loop in the close conformation and DFG-flip in the "out" conformation, PDB ID: 1IEP). The smaller amino-terminal part is the upper lobe, while the lower lobe represents the carboxyl-terminal part. The catalytic site lies in a cleft between the two lobes, and the relative orientations of the two lobes open or close the cleft. The activation loop adopts different conformations in the active/inactive states and it begins with the DFG sequence. (Bottom) Focus on the DFG motif (left) in the "in" conformation and (right) in the "out" conformation.

Riboswitch-ligand

For the PreQ1 riboswitch bound to its two ligands, the topology and initial structure are derived from a recent study by Wang *et al.* [238]. Here, the RNA-ligand complexes are modeled using the X-ray crystallographic structures with PDB codes 6E1W and 6E1U for the complexes with the cognate and synthetic ligands, respectively. Ligands are parametrized using the GAFF and AM1-BCC charges, accounting for a total charge of +1 due to the presence of a quaternary nitrogen in both ligands. TIP4P-D water model [244] is used to solvate the systems. This water model is a a four-site water model commonly employed in combination with the DESRES FF for RNA simulations [245]. To neutralize the systems, sodium and chloride ions are added to achieve a salt concentration of 0.15 M. The solvated systems are then equilibrated in two steps, following the procedure outlined in Reference [238]. Additionally, the cognate ligand, is re-parameterized using a more accurate charge representation by deriving RESP charges [75] and optimizing dihedral parameters. RESP charges are computed using QM methods with the 6-311++G** basis set at the wB97X-D level of theory, while the dihedral angles are optimized using the xTB method [246]. Because of the starting X-ray structure (PDB: 6E1W) lacks hydrogen atoms, the cognate ligand may adopt two different tautomeric states when bound to the RNA receptor. To account for this, an additional complex is constructed, starting from the X-ray structure, but with a different protonation state for the ligand, representing the alternative tautomeric form.

6.4 Results and Discussion

In this section, we discuss the results for these two complex systems, characterized by their large size and significant conformational flexibility.

6.4.1 Abl-Gleevec: ligand flexibility and large conformational rearrangements

The Abl-tyrosine kinase complexed with its inhibitor Gleevec is significant both therapeutically and structurally. Gleevec is a widely used cancer drug with strong inhibitory potency, experimentally shown to bind with a ΔF_b of -10.9 kcal/mol [209].

In our previous study (Chapter 4 [P1]), we faced significant challenges with this system due to its large size and intricate binding mechanism. In this work, we successfully address these difficulties, achieving a reliable binding free energy estimate for the Abl-Gleevec complex. Furthermore, we optimize the computational workflow to extend our protocol applicability to larger, pharmaceutically relevant biomolecular systems with intricate structures and dynamics, using Abl-Gleevec as a representative prototype.

Path definition and PCVs for Abl-Gleevec

As discussed in Chapter 5 [P2], proper parametrization of PCVs is essential for minimizing the dissipated work generated during SMD simulations. This is a critical factor, as excessive dissipation (i.e., heat production) can significantly hinder the convergence of binding free energy estimates. Improved PCVs can be achieved by incorporating, alongside ligand atoms, additional system degrees of freedom that are relevant to the binding process into the reference path. This primarily affects the definition of the $S(x)$ PCV, as these added components are pulled by the SMD bias during the simulation.

To improve the PCV definition for the Abl-Gleevec system, we visually analyze the binding trajectories, as the binding process (rather than unbinding) is the primary source of dissipated work. The simulations show that the rapid pulling of the ligand into the binding pocket do not allow adequate time for the pocket to undergo the necessary conformational adjustments. Consequently, the Abl binding site remains largely closed as the ligand approaches, resulting in substantial dissipated work. This analysis is consistent with previous studies [221, 247, 248].

Here, the ligand binding to the apo-kinase is described as a highly complex process involving a hybrid mechanism of conformational selection and induced fit. Additionally, significant conformational changes in the activation loop and glycine-rich loop must be considered, as they enable the Abl binding pocket to adopt the distinctive tunnel-like shape previously reported [243]. For this reason, a key degree of freedom in the overall process is the opening and closing of the Abl binding pocket during both binding and unbinding simulations. To accurately capture this conformational rearrangement in SMD simulations, heavy atoms of the binding pocket located within 6 Å of the ligand in its bound pose are included in the reference path. Therefore, during binding SMD simulations, the bias potential pulls the ligand into the binding pocket and also guides conformational rearrangements within the protein binding site. This improved reference path is expected to reduce dissipated work and to facilitate the structural changes required for effective binding mechanism.

Moreover, an optimal path definition also requires that the ligand exit trajectory (particularly its main axis) remains as orthogonal as possible to the protein surface during unbinding. This consideration is particularly critical for proteins like Abl, which feature solvent-exposed binding pockets. Furthermore, traditional solvation shell definitions, suitable for smaller ligands, fail to adequately characterize the unbound state of large and charged ligands like Gleevec. In the final configuration of the reference path, the ligand must be placed at least 15 Å from the protein surface to minimize residual electrostatic interactions in the unbound state. This value corresponds to more than three solvation shells. Finally, the reference path must also capture all relevant intermediate metastable states that occur during the binding process.

As previously discussed, the initial trajectories used to define PCVs are generated through exploratory studies involving multiple ABMD simulations. Longer ABMD runs can promote a more gradual, realistic, and spontaneous unbinding process. Therefore, using the same setup described in Chapter 4, we perform several 20 ns ABMD simulations (increased from the previous 10 ns). These simulations are designed to sample the unbinding of Gleevec from Abl, lasting until

a separation of at least 15 Å is reached between the centers of mass of the ligand and the binding pocket. Based on the observed frequency of the unbinding mechanism and the orthogonality of the ligand trajectory relative to the protein surface, we select a candidate guess path from this ensemble of simulations. This path is then refined using the path optimization algorithms, resulting in a refined pathway composed of 42 equidistant molecular configurations. Each configuration includes the ligand and the heavy atoms of binding pocket residues located within 6 Å of the ligand in the bound pose, totaling 51 residues. Among these residues (used to refine the path definition) we ensure the inclusion of residues known to be involved in key protein conformational changes, as supported by both literature reports and our ABMD trajectories. Specifically, this selection encompasses the DFG motif (Asp381, Phe382, Gly383), the initial residue of the activation loop (Leu384), and most residues from the glycine-rich loop. Notably, two highly conserved residues from the glycine-rich loop, Leu248 and Val256, are incorporated, as they consistently form interactions with Gleevec across bound states and remain in contact throughout the unbinding process until the ligand becomes fully solvated. Figure 6.2 depicts the key structural configurations that define the refined reference path. This path begins with the X-ray structure of the Abl-Gleevec complex, where Gleevec is in its bound pose. In this conformation, the pyridine and pyrimidine rings of the inhibitor occupy the same region as the natural substrate, while the rest of the molecule extends into the hydrophobic core of the kinase, effectively stabilizing it in an inactive state. Along this reference pathway, substantial conformational rearrangements in the P-loop and C-helix facilitate the unbinding of Gleevec, accompanied by coordinated movements of the activation loop. The ligand exits orthogonally through the cleft between the N-terminal and C-terminal lobes, reaching a fully unbound state in the solvent. These structural transitions are consistent with previously reported mechanisms of Gleevec binding to Abl [221, 219]. However, fully capturing the conformational transition of Abl from its inactive to active state remains challenging and would require plain MD simulations on the microsecond timescale [221]. Consequently, the reference path derived from fast ABMD simulations does not capture the DFG-motif flip within the activation loop, nor does it reflect the full extent of its conformational rear-

range, which includes a rotation of the A-loop and displacements of up to 35 Å for central residues [219]. It is important to note that our aim is to apply a standard protocol rather than tailoring the protocol for each specific system. As a result, the activation process is not fully captured, and the generated pathway provides only a simplified, coarse-grained representation of the binding event. Finally, to reduce boundary artifacts in the PCVs that could distort the free energy surface, we append fictitious configurations (generated through short SMD runs) at both ends of the reference path. Although these configurations are not physically sampled during the actual simulations, they enhance the representation of the true bound and unbound states, thereby improving the overall quality of the sampling.

Binding free energy estimates

Using the newly refined reference path, composed of 42 molecular configurations, we conduct 30 SMD production runs for both the binding and unbinding processes. Simulations are carried out at two different durations (i.e., 100 ns and 200 ns) corresponding to different pulling velocities, as a constant-velocity restraint is applied. To evaluate whether these simulation times are sufficient to achieve convergence of the free energy estimates, we conduct a statistical analysis assessing convergence as a function of the number of replicas.

The Jarzynski work profiles are calculated over simulation times of 100 and 200 ns for both binding and unbinding processes. These profiles are shown in Figure 6.4, with the upper panel depicting the binding trajectories and the lower panel showing the unbinding ones. These profiles reveal a notable reduction in dissipated work during the SMD simulations compared to those presented earlier in Section 4 attributable to the inclusion of pocket atoms in the reference path. Additionally, the close similarity between the work curves at 100 ns and 200 ns indicates that a 100 ns simulation time already achieves a sufficiently gentle pulling speed to approach convergence.

Using our CFT-based approach, we reconstruct the free energy profiles along $S(x)$, as shown in Figure 6.5. Furthermore, Figure 6.8 provides detailed bootstrap error estimates for each point on the free energy profiles. As shown in these plots,

the free energy profile at 100 ns exhibits a higher activation energy barrier for binding. This is likely because faster SMD simulations are more influenced by kinetic effects, which can lead to artificially elevated barriers in the free energy landscape. However, the similarity between these free energy profiles (and the work curves) at both pulling speeds means that 100 ns is adequate for obtaining thermodynamically accurate, though not kinetically detailed, results.

To calculate the binding free energy, we compute the ratio of the bound and unbound partition functions, using the 18th configuration as discriminating structure between these states. After applying the standard volume correction, the resulting standard binding free energies are found to be in agreement with the experimental value of -10.9 kcal/mol. Particularly, the CFT-derived binding free energy estimates from the 100 ns and 200 ns SMD simulations are -12.5 ± 1 kcal/mol and -13.0 ± 2 kcal/mol, respectively. These results demonstrate that incorporating the protein pocket degrees of freedom reduces dissipated work, thereby enhancing the convergence and precision of the binding free energy calculations

Finally, we assess the necessity and impact of the new path criteria by applying an independent sampling method. Specifically, we run WT-MetaD simulations using both the previous and the newly refined reference paths. With the refined path, Gleevec is able to explore regions further from the protein, more thoroughly sampling the bulk unbound state. The resulting free energy and $S(x)$ profiles are shown in Figures 6.7, 6.6. These plots demonstrate that Gleevec is able to move away from the protein surface, allowing an exhaustive sampling of the unbound state. This qualitative improvement highlights the significance of the newly introduced path refinement criteria.

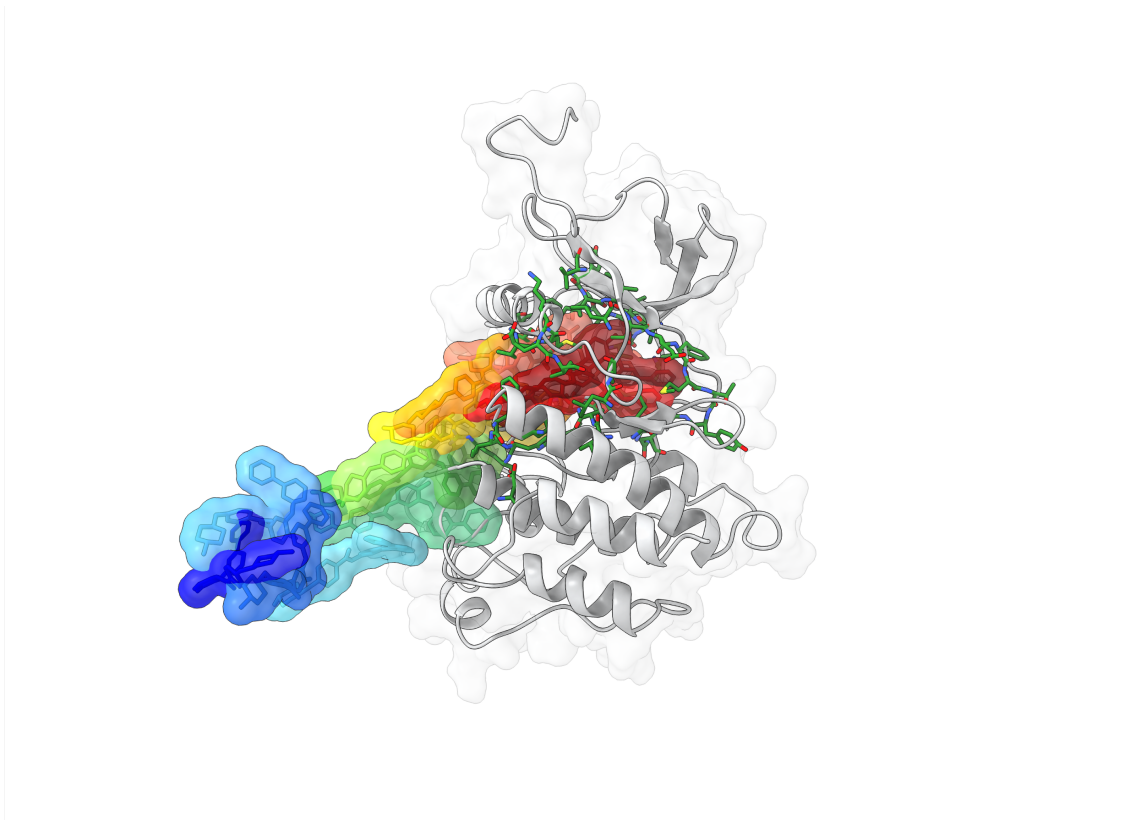


Figure 6.2: Schematic representation of the Abl-Gleevec complex's path. Abl is in grey in a ribbon representation, with the pocket atoms included in the path in green and licorice. The surfaces of the path configurations of Gleevec are represented from dark red for the bound state to dark blue for the unbound state.

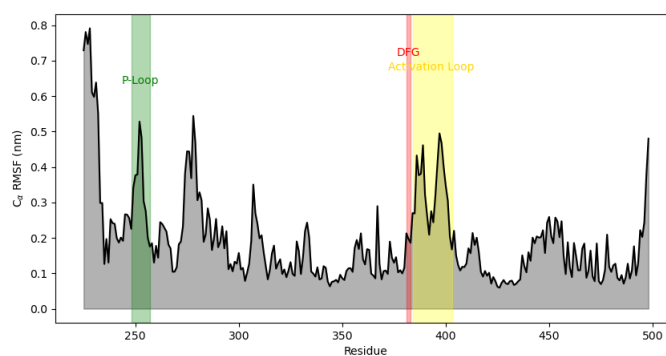


Figure 6.3: Fluctuations of the P-loop (green) and of the A-loop (yellow), with highlighted in red the DFG sequence. Conformational rearrangements occur for the loops during simulations, while the DFG sequence remains more stable

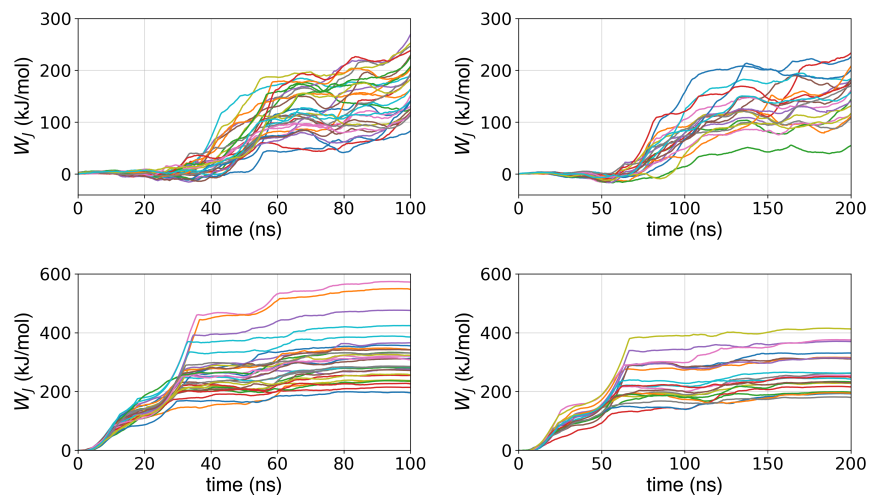


Figure 6.4: Jarzynski work profiles of Abl-Gleevec for binding (upper section) and unbinding simulations (lower section), measured over simulation times of 100 and 200 ns

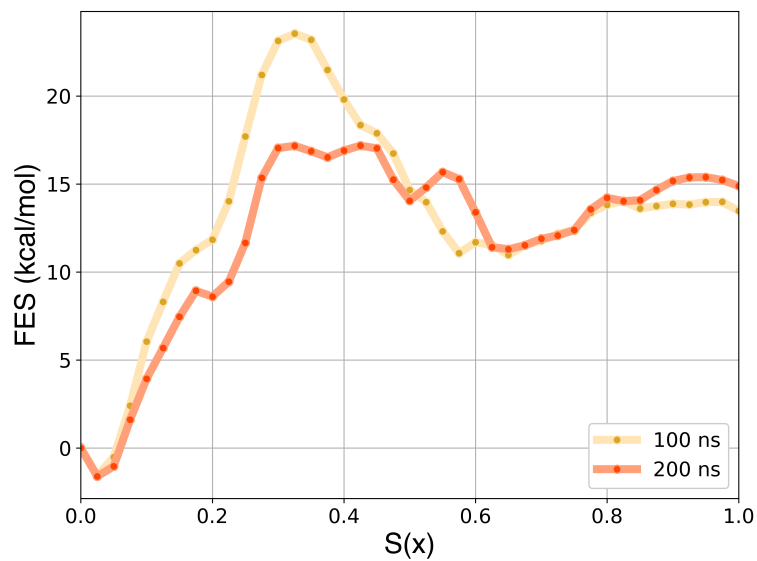


Figure 6.5: Free energy profiles along $S(x)$ obtained by applying CFT to SMD of 100 and 200 ns with the refined PCVs for Abl-Gleevec.

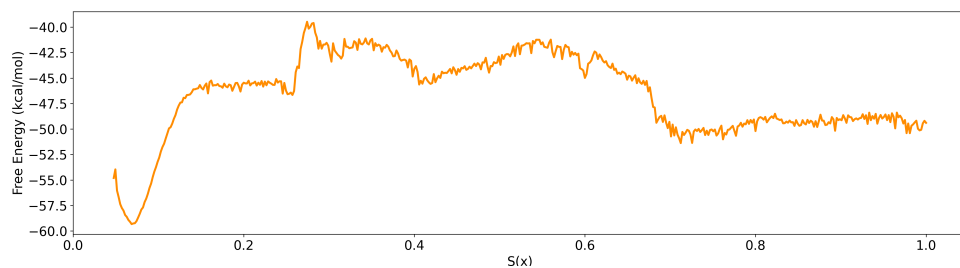


Figure 6.6: Reweighted free energy profile as a function of the collective variable $S(x)$ from metadynamics simulation for Abl-Gleevec system.

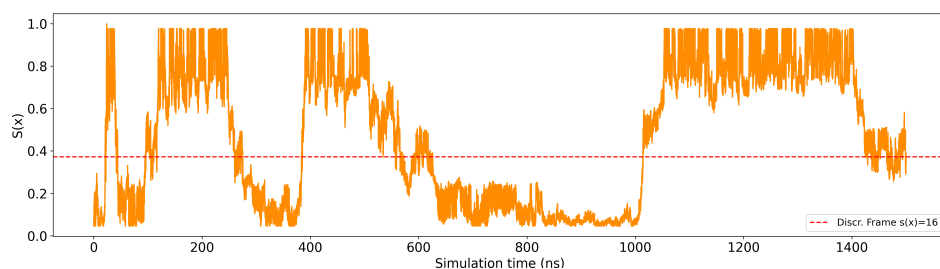


Figure 6.7: Time series of the collective variable $S(x)$ along the metadynamics trajectories for Abl-Gleevec system. The red dotted line represents the discriminating frame used for free energy calculations

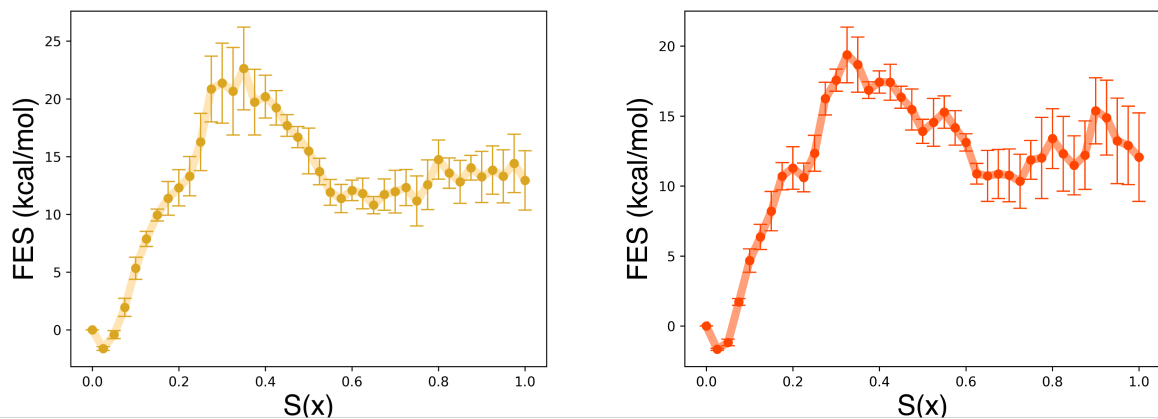


Figure 6.8: Free energy profiles along $S(x)$ obtained by applying CFT to SMD of 100 and 200 ns with the refined PCVs for Abl-Gleevec, along with their corresponding bootstrap errors.

Conformational rearrangements and DFG-Flip transition

The key role of the DFG-motif in mediating Gleevec inhibition of Abl has been demonstrated through numerous experimental and computational studies [243, 221, 249]. Schindler *et al.* [250] report that Gleevec binding to the Abl pocket requires a conformational change in the DFG-motif, involving a 180° rotation known as the DFG-flip, which transitions the motif from an active to an inactive state. In the Abl-Gleevec complex, the DFG-motif adopts the out inactive conformation, while the activation loop remains closed, thereby preventing the binding of the natural substrate. Additionally, the glycine-rich loop folds into a kinked conformation directed toward the binding pocket, promoting favorable interactions with the bound ligand. This conformation of the glycine-rich loop is stabilized by a specific hydrogen bond between Tyr253 and Asn322. Several computational studies have been conducted to evaluate the impact of the DFG-flip on the overall binding free energy of the Abl-Gleevec complex. For instance, Roux and colleagues [243] report that the DFG-flip contributes significantly to binding, accounting for approximately 1.4 kcal/mol of the total free energy.

However, throughout our 200 ns simulations, the DFG-flip is not observed. While conformational changes occurred in both the P-loop and A-loop, the DFG-motif remains comparatively stable. This behavior is illustrated in Figure 6.3, where the DFG sequence is highlighted in red. This RMSF analysis further confirms the dynamic flexibility of the loops compared to the relative stability of the DFG motif in its out conformation. Even if our refined reference path includes the DFG-motif residues, the ABMD simulations did not capture the transition of this motif from the closed to open state. Simulating the DFG-flip would require either a specifically designed collective variable or significantly longer simulation times. For example, Lovera *et al.* [251] observe this rearrangement only after running a 200 μ s plain MD simulation. Accordingly, during our 20 ns unbinding ABMD runs, using an electrostatic-like collective variable between ligand and pocket atoms, the DFG motif remains in the out conformation even as the ligand reaches the unbound state.

To qualitatively account for the significant free energy contribution of the DFG-flip

into our results, we apply a correction term of $\Delta F_{in \rightarrow out} = 1.4$ kcal/mol (as reported in Reference [243]) to our standard binding free energy estimates. Specifically, the standard binding free energy of the Abl-Gleevec complex, $\Delta F_{bind,in \rightarrow out}^\circ$ corresponds to the following reaction:



However, we compute $\Delta F_{bind,out}^\circ$ corresponding to a reaction where the DFG-motif of Abl is in the out conformation in both the bound and unbound states:



Therefore, by adding $\Delta F_{in \rightarrow out}$ to $\Delta F_{bind,out}^\circ$, we can recover the standard binding free energy accounting for the DFG-flip:

$$\Delta F_{bind,in \rightarrow out}^\circ = \Delta F_{bind,out}^\circ + \Delta F_{in \rightarrow out} \quad (6.3)$$

Including this contribution, the CFT estimates for the 100 ns and 200 ns simulations adjust to -11.1 and -11.6 kcal/mol, respectively, thus bringing the results closer to the experimental value.

6.4.2 Riboswitch-ligand: Receptor Flexibility and Conformational Rearrangements

Building on the methodology developed for protein systems, we extend our investigation to a challenging class of biomolecules: RNAs.

Path definition and PCVs for RNA-ligand systems

As shown, properly parametrized PCVs reduce dissipated work in nonequilibrium SMD simulations, enhancing the convergence of free energy estimates. Likewise, establishing an optimized and realistic unbinding pathway is equally crucial for RNA-ligand systems.

We perform 10 exploratory ABMD simulations, each lasting 10 ns, to generate initial trajectories for the two RNA-ligand complexes in the TIP4P-D water model. These simulations are guided by a CV based on the Debye-Hückel interaction energy, capturing the electrostatic interactions between the highly negative RNA charges and the positively charged ligands. Specifically, these simulations are guided toward a CV value of zero, representing complete ligand unbinding. Trajectories are truncated once the ligand-RNA distance exceeded 20 Å to ensure a realistic unbound state. As shown in Figure 6.9, during these runs both ligands consistently dissociate via a solvent-exposed pathway situated between loop 2 (residues U12-U13-A14-U15-A16-C17) and stem 1 (residues U9-A10-G11 and C31-U32-A33-A34). This recurring unbinding route involves structural rearrangements of loop 2 and stem 2, which opens the binding pocket and facilitates ligand release. The 3D structure of the riboswitch depicted in Figure 6.9 emphasizes its key structural elements, specifically:

- Loop 1: residues U6-U7-C8;
- Loop 2: residues U12-U13-A14-U15-A16-C17;
- Loop 3: residues A23-U24-A25-A26-A27-A28-A29-A30;
- Stem 1: residues A1-G2-A3-G4-G5 and C18-C19-U20-C21-U22;
- Stem 2: residues U9-A10-G11 and C31-U32-A33-A34.

Figure 9.5 in the Supporting Information section (Chapter 9) shows the Debye-Hückel interaction energy throughout these ABMD simulations. The graphs illustrate a clear decrease in interaction energy as the ligands dissociate, reaching final values of 1.73 kcal/mol for the cognate ligand and 1.14 kcal/mol for the synthetic ligand.

A major difference between protein and RNA systems is the significantly higher structural flexibility of RNA [252, 232, 231, 253]. As a result, establishing a reference pathway for PCVs in RNA-ligand complexes required tailored modifications to accommodate this increased dynamism. The PCV calculation assumes that the optimal-alignment RMSD between the reference path configurations and the

instantaneous simulation configuration is inherently reliable. For protein-ligand complexes, alignment is typically performed using the protein’s $C\alpha$ atoms. However, due to the pronounced flexibility of RNA, selecting appropriate atoms for alignment is more challenging but essential for obtaining accurate PCV values and alignment. To address this, we conduct three independent 100 ns unbiased MD simulations and combined the data to compute residue-wise RMSF values, as shown in Figure 9.6 in SI. Residues with RMSF values below $1.8 \text{ \AA}$ are considered stable, and from these, we select specific atoms for alignment: the phosphorus (P) and $C1'$ carbon atoms of the RNA backbone (see Figure 9.7 in SI). This choice aligns conceptually with the approach used for proteins, ensuring a consistent basis for PCV calculation. Moreover, selecting the appropriate RNA and ligand atoms to include in the reference pathway is particularly critical for RNA-ligand systems. This is because ligand (un)binding events often involve significant conformational changes in regions such as stem 1, which forms a complex hydrogen-bond network with the cognate ligand. Notably, structural adaptations, including specific conformations of residue C15, are necessary to accommodate the bulkier synthetic ligand in the binding site (Figure 6.10) [239]. Thus, to accurately capture these structural changes, atoms within $6 \text{ \AA}$ of the ligand during ABMD unbinding trajectories are incorporated into the path definition. For purine nucleotides, the atoms $N1$, $C8$, and $C1'$ are selected, while for pyrimidines, $C6$, $N3$, and $C1'$ are used (see Figure 9.7 in SI). This atom selection strategy results in a refined unbinding reference path composed of approximately 35 equidistant configurations for each RNA-ligand system, as illustrated in Figure 6.11.

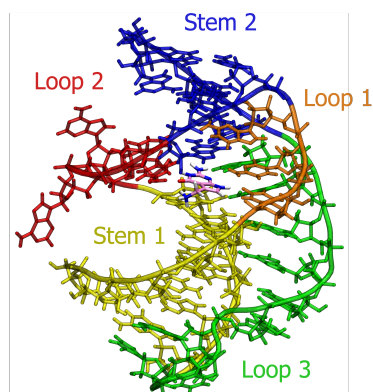


Figure 6.9: 3D structure of Riboswitch preQ1 with its cognate ligand (Q1) with residues colored according to their secondary structure.

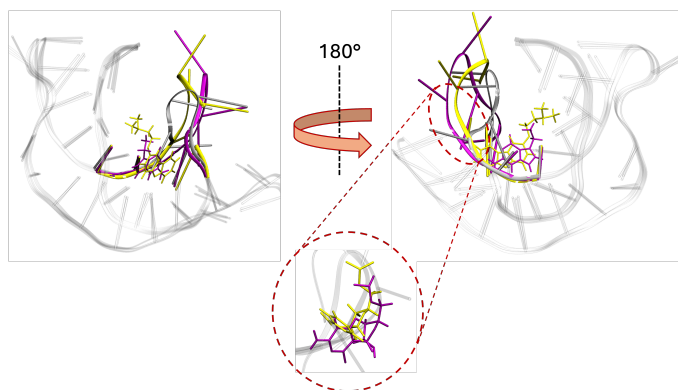


Figure 6.10: Structural superposition of the apo (white), cognate ligand-bound (purple), and synthetic ligand-bound (yellow) states of the receptor, highlighting the conformational rearrangement of loop 1 (distinct orientations of residue C15).

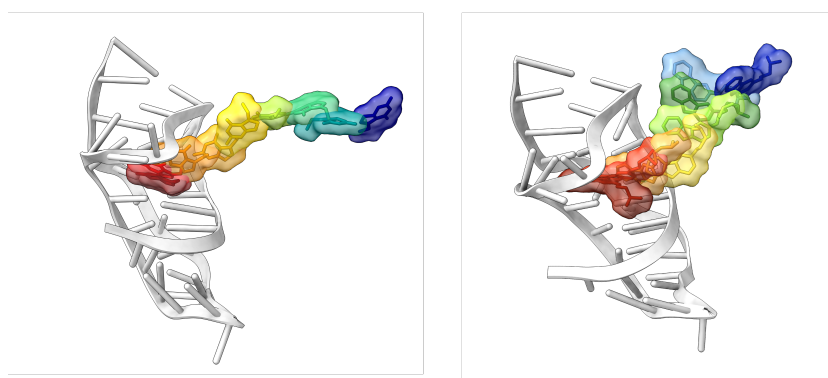


Figure 6.11: Schematic representation of the paths of riboswitch-preQ1 in complex with (left) the cognate ligand and (right) the synthetic ligand. The RNA is shown in light gray.

Binding free energy estimates

Following our established protocol, we conduct 30 replicates of bidirectional SMD simulations, each lasting 100 ns. The Jarzynski work is computed for each trajectory, and the resulting work profiles over time are shown in Figure 6.12. Figure 6.12A presents the unbinding work profiles for the cognate ligand, which increases almost monotonically. This indicates the presence of an energy barrier that needs to be overcome by the ligand to disrupt the interactions of the bound state and escape the binding pocket. These simulations reveal greater work dissipation associated with breaking the hydrogen-bonding network within the binding pocket, in contrast to the synthetic ligand. Once the ligand reaches a fully solvated unbound state, the work values stabilize (plateau at around 45 ns), indicating the completion of the unbinding process. Conversely, Figure 6.12B shows the work profiles for the binding simulations, where the ligand initially resides in the solvent, resulting in near-zero work values. As the ligand approaches the RNA target (around 45 ns into the simulation) the work begins to rise, reflecting the energetic cost of entering the binding site. This increase continues until the bound state is achieved. Overall, the work required to reach the bound state appears comparable for both the cognate and synthetic ligands. The dissipation variability across different replicas highlights the intrinsic complexity of the binding process. Despite this, the overall work dissipation remains modest in both cases, indicating that the reference path criteria effectively capture the key structural transitions and are well-suited for modeling RNA-ligand interactions.

Using our protocol, we reconstruct the free energy profile along $S(x)$ for both the cognate and synthetic ligands in the TIP4P-D water model, as shown in Figure 6.13. To estimate the binding free energy we select the discriminative conformations corresponding to $S(x)$ values of 16 for the cognate ligand and 14 for the synthetic one. The estimated standard binding free energies, which include the standard volume correction, are compared with experimental affinity data reported in recent studies [238, 239]. For the RNA-synthetic ligand complex, the calculated ΔF° is -5.6 ± 1 kcal/mol, showing a modest deviation from the experimental reference of -7.9 kcal/mol. In contrast, the RNA-cognate ligand complex yielded an estimated ΔF°

of -17.2 ± 1 kcal/mol, indicating a more substantial deviation from the experimental value of -10.9 kcal/mol.

While CFT-derived free energy profiles for the cognate ligand appears consistent, the profiles derived from binding simulations using the JE display a scattered and irregular pattern, as depicted in Figure 6.14A. This behavior may stem from the reduced ability of JE to compensate for dissipative work during highly nonequilibrium processes. Dissipation likely contributes to the deviation observed in the binding free energy estimate for the cognate ligand.

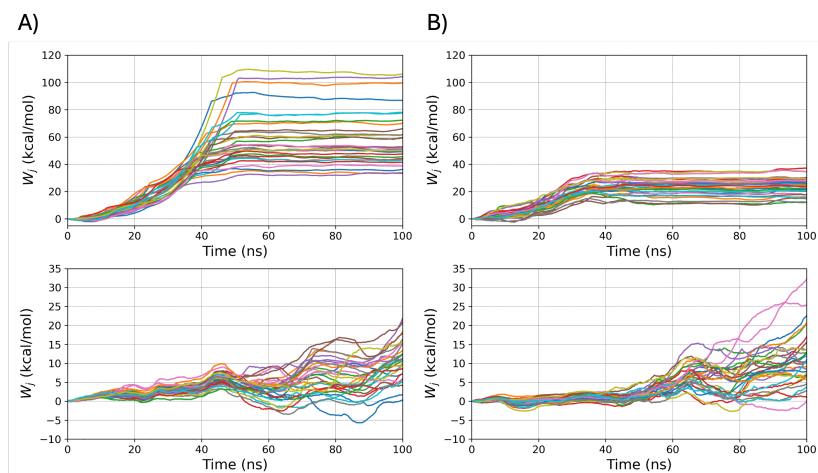


Figure 6.12: Work profiles for the cognate ligand (A) and synthetic ligand (B) regarding unbinding (upper section) and binding (lower section) 100 ns simulations in TIP4P-D.

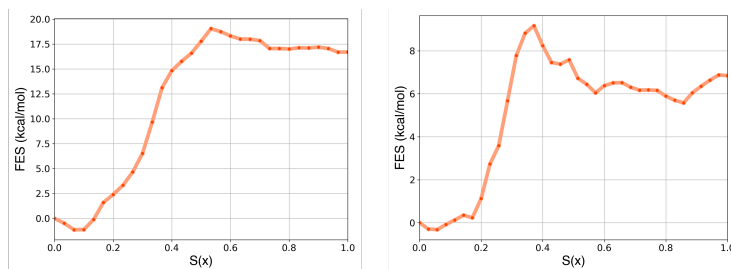


Figure 6.13: Free energy profiles along $S(x)$ obtained by applying CFT to SMD (simulation time of 100 ns in TIP4P-D) for the cognate (A) and synthetic (B) ligands

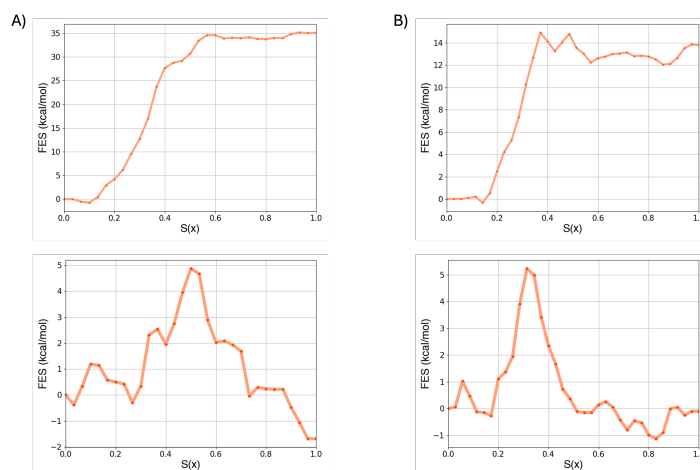


Figure 6.14: JE-derived FES for unbinding (upper section) and binding (lower section) simulations in TIP4P-D for the cognate (A) and synthetic (B) ligands.

Water model effect on binding free energy estimates

Up to this point, all simulations have been conducted using the TIP4P-D water model. However, as discussed in Section 5 [P2], four-site models may not provide the best agreement with experimental data in the context of nonequilibrium simulations. Indeed, the TIP3P water model may exhibit reduced sensitivity to dissipative effects, potentially resulting in free energy estimates with lower bias. While TIP3P is neither the most advanced nor the most accurate model for representing RNA hydration, its computational efficiency and lower dissipation make it a pragmatic choice for nonequilibrium simulations.

Therefore, in line with the rationale outlined in Section 5, we investigate the potential impact of the water model on binding free energy predictions. To this end, we repeat the entire set of SMD simulations using the TIP3P water model, which is expected to introduce less dissipative work. Importantly, all other parameters and setup, including the reference path, are kept unchanged to isolate the effect of the solvent model. The work profiles obtained with the TIP3P water model are shown in Figure 6.15. For the synthetic ligand (Figure 6.15B), the profiles resemble those from TIP4P-D simulations (Figure 6.12B). In contrast, for the cognate ligand, the choice of water model has a more pronounced impact, as reflected in the work curves (Figure 6.15A) and the resulting free energy values summarized in Figure 6.16. Based on our results, the most marked differences between the TIP3P and TIP4P-D water models emerge during the binding simulations. Similar to what was observed for protein-ligand systems, the TIP3P model exhibits reduced dissipation in RNA-ligand systems as well. This results in a lower amount of work required to achieve the bound state compared to TIP4P-D, as illustrated in Figure 6.12A.

The visual analysis of the MD simulations reveals notable conformational differences in the RNA bound state when comparing results obtained with TIP3P and TIP4P-D water models. Specifically, simulations using TIP4P-D show difficulties in accurately following the reference pathway during the final stages of binding. In these cases, stem 1 fails to adopt the precise conformation observed in the X-ray structure, indicating limited structural convergence near the bound state.

As a result, for the synthetic ligand, the CFT-derived profile (Figure 6.17B) and the corresponding binding free energy obtained using the TIP3P water model differ from those calculated with TIP4P-D. Specifically, the estimated standard binding free energy with TIP3P is -8.7 ± 0.7 kcal/mol (14th configuration of the reference path selected as discriminating frame), which shows improved agreement with the experimental value. In contrast, the TIP4P-D model produced a less accurate estimate of -5.6 ± 1 kcal/mol. For the cognate ligand, both the free energy profile (Figure 6.16A) and the standard binding free energy estimated using the TIP3P model are comparable to those obtained with TIP4P-D. The TIP3P-based estimate of ΔF° is -18.5 ± 3 kcal/mol, which, despite the consistency across water models, still significantly deviates from the experimental reference value of -10.9 kcal/mol.

The findings for the synthetic ligand underscore the critical role of the water model in determining the accuracy of nonequilibrium free energy calculations. Simulations performed with the TIP4P-D model tended to underestimate the binding free energy, whereas those using the TIP3P model produced estimates more consistent with experimental measurements, falling within the expected error margins. However, even when using the less viscous TIP3P model, the SMD simulations for the cognate ligand continue to show a substantial deviation from the experimental binding free energy values.

To deeper understand the impact of the nonequilibrium approach on RNA-ligand systems, we complement our analysis with an independent sampling method. Specifically, we carry out WT-MetaD simulations using the previously defined PCVs and reference pathways, along with the TIP4P-D water model. These simulations enables thorough sampling along the $S(x)$ PCV, successfully capturing multiple binding and unbinding transitions for both ligands within a single simulation (SI of Reference [P3]). This outcomes indicate that the employed PCVs and reference paths are capable of describing the essential conformational space involved in the ligand (un)binding processes. Notably, the cognate ligand remains predominantly in the bound state throughout a large portion of the MetaD simulations. This behavior can be attributed to its strong interaction network with the RNA in the bound conformation. Following the reconstruction of the free energy profile,

	SMD (TIP4P-D)	SMD (TIP3P)	MetaD
Cognate	-17.2 ± 1	-18.6 ± 3	-19.4 ± 2
Synthetic	-5.6 ± 1	-8.7 ± 0.7	-7.4 ± 0.9

Table 6.1: Standard binding free energies (kcal/mol) for the cognate and synthetic ligands using the TIP4P-D water model in MetaD simulations and TIP4P-D and TIP3P in Steered Molecular Dynamics.

the corresponding binding free energies are computed and they are summarized in Table 6.1.

The binding free energies obtained from WT-MetaD differ from those calculated using SMD with the TIP4P-D water model. This comparison is rather interesting because WT-MetaD does not rely on the Jarzynski work for the free energy estimation. On the other hand, the nonequilibrium approaches based on the CFT and JE estimators are heavily dependent on this parameter. Interestingly, the MetaD results using TIP4P-D align more closely with the SMD results performed in TIP3P water, indicating that TIP3P may be more appropriate to reduce dissipation. This characteristic can make it a more suitable choice for nonequilibrium SMD simulations. Conversely, computationally more demanding water models like TIP4P-D may offer enhanced accuracy for equilibrium-based methods, such as MetaD.

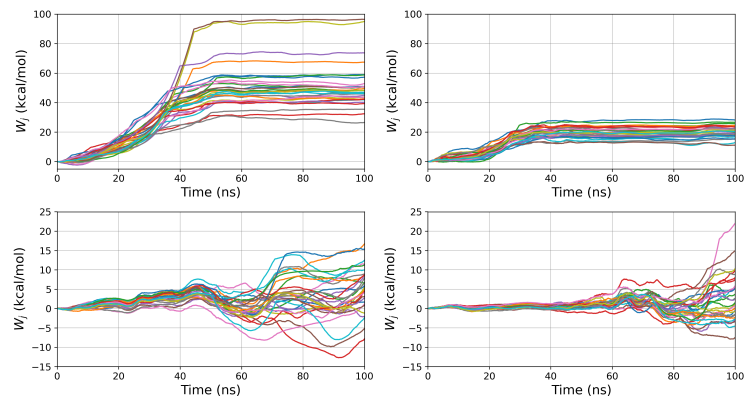


Figure 6.15: Work profiles for the cognate ligand (A) and synthetic ligand (B) for unbinding (upper section) and binding (lower section) 100 ns simulations in TIP3P.

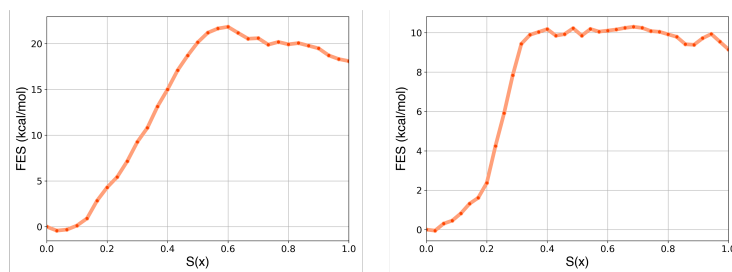


Figure 6.16: FCFT-derived FES along $S(x)$ (simulation time of 100 ns in TIP3P) for the cognate (A) and synthetic (B) ligands

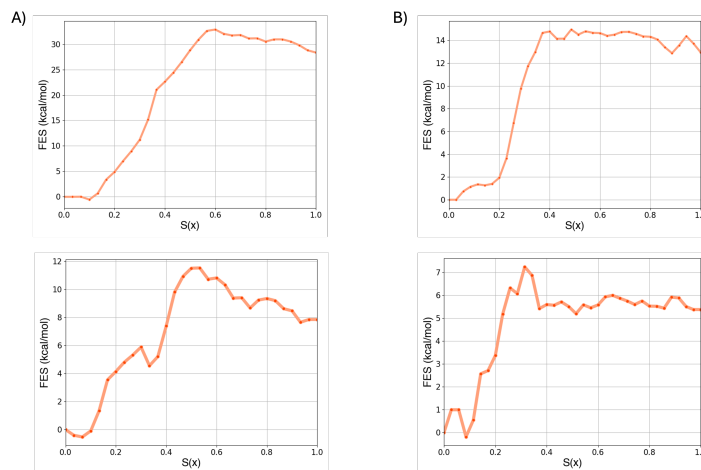


Figure 6.17: JE-derived FES for unbinding (upper section) and binding (lower section) simulations in TIP3P for the cognate (A) and synthetic (B) ligands.

Possible source of errors for the cognate ligand

The binding free energy values calculated for the cognate ligand show a significant deviation from the experimental reference. Both SMD and WT-MetaD simulations suggest a stronger binding affinity for the cognate ligand than what is observed experimentally. In contrast, the binding affinity estimated for the synthetic ligand aligns well with the experimental data. The experimentally observed higher affinity of the cognate ligand likely contributes to this discrepancy. Because the cognate ligand binds more tightly to the RNA target, accurately estimating its binding affinity is inherently more challenging than for the synthetic ligand.

Particularly, Figure 6.20B and Figure 6.18 illustrate the crystallographic binding poses of both ligands, highlighting that the cognate ligand engages in a greater number of interactions with the RNA target, which drives its binding mode. Both ligands exhibit a similar extent of stacking interactions with the RNA. However, the cognate ligand forms seven hydrogen bonds with the target, in contrast to just two hydrogen bonds formed by the synthetic one. Thus, the strong interactions between the RNA and the cognate ligand introduced challenges in achieving an accurate estimation of the binding free energy for this complex.

In this context, we examine potential factors contributing to the observed discrepancies between our calculated results for the cognate ligand and the corresponding experimental binding affinity. One possible source of error may stem from limitations in the FFs used for RNA-ligand simulations. Although FFs for macromolecules like proteins and RNAs have seen significant advancements and now offer increasingly reliable and quantitatively accurate predictions of experimental data, small molecule parameterization still remains problematic. As explained in Section 2.3.2, this difficulty mainly arises from the wide diversity and complexity of the chemical space occupied by small organic compounds, which makes accurate modeling more demanding. Particularly, inaccuracies in parameters such as atomic charges and dihedral angles can substantially affect simulation outcomes [254]. Currently, it is common practice to use the AM1-BCC method for assigning partial charges to ligands in simulations, as it generally provides a reliable description of ligand electrostatics. However, in more complex cases, AM1-BCC charges may not

be sufficiently accurate. Under such circumstances, higher-level approaches, such as deriving charges from Density Functional Theory calculations combined with the RESP fitting method, can offer improved accuracy [75]. For the systems examined here, many heteroatoms of the ligand form critical hydrogen bonds with the RNA target in the bound state. To better capture these interactions, we refine the parametrization of the cognate ligand by adopting RESP-derived charges and optimizing dihedral angle parameters. Next, we apply our protocol to produce a new initial unbinding pathway using ABMD simulations, followed by production SMD runs. Notably, the preferred unbinding route, passing between loop 2 and stem 1, remains consistent with that observed using the previous ligand parametrization and results reproducible across all ABMD trajectories. To evaluate the influence of different solvent models combined with the updated ligand parameters, simulations are performed with both TIP4P-D and TIP3P water models. Table 6.2 summarizes the outcomes of these simulations.

	TIP4P-D	TIP3P
N1 - RESP	-21.2 ± 1	-21.4 ± 1
N1 -AM1-BCC	-17.2 ± 1	-18.6 ± 3
N2 -AM1-BCC	-13.2 ± 1	-11.9 ± 1

Table 6.2: Binding free energies (kcal/mol) for the cognate ligand in the tautomeric form N1 with RESP and AM1-BCC charges and the tautomeric form N2 with AM1-BCC charges in different water models.

According to these results, the binding free energy values show minimal sensitivity to the choice of water model after applying RESP parametrization. The estimates obtained using TIP3P and TIP4P-D solvents are comparable and fall within the statistical error, consistent with the trend observed when using AM1-BCC charges. However, the results obtained with RESP parameters resemble those with AM1-BCC charges, with both sets deviating from the experimental values. This suggests that the choice of charge model may not be the primary source of error in this case.

To further validate our findings and conclusively exclude ligand parameterization as a source of error, we reparameterized all ligands, including both the synthetic and cognate compounds. Instead of relying on the topologies from Reference [238],

we rebuild the ligand topologies (i.e the synthetic and the cognate one in both tautomeric forms N1 and N2) using the AM1-BCC charge model combined with the GAFF. Additional simulations (with both TIP4P-D and TIP3P water models) are performed for all complexes using this newly generated set of parameters. With these updated parameters, we repeat the entire workflow, (i.e., ABMD, path definition, two-sided SMD simulations) obtaining the binding free energy values in the end. Figure 6.19 summarizes all these additional results. The consistency of these outcomes with previous findings provides further evidence that parameter selection is not the primary source of error for the cognate ligand. Moreover, it demonstrates that our nonequilibrium pipeline is both reproducible and stable.

The second source of inaccuracy for the cognate ligand may arise from the presence of different tautomeric states in solution.[255]. We conduct QM calculations using Jaguar, a Schrödinger software package, to investigate the two tautomeric forms of the cognate compounds, referred to as N1 and N2. These tautomers differ by the position of a hydrogen atom on one of the nitrogen atoms in the pyrimidine ring. Specifically, the previously considered cognate tautomer, referred to as N1, can undergo tautomerization through the migration of a proton from the N1 nitrogen to the N2 nitrogen atom, resulting in the alternative tautomer N2, as illustrated in Figure 6.20A. In solution, the N2 form may be stabilized by an intramolecular hydrogen bond formed between the carbonyl group and the protonated amine, potentially favoring its population. To assess the relative stability and potential coexistence of these forms, we employ ab initio QM calculations with the B3LYP-D3 functional and the 6-311G++** basis set. Solvation effects are accounted for using the Poisson-Boltzmann Finite Element method. Our results indicate that the hydration free energy of the N1 tautomer is favored by 1.85 kcal/mol compared to N2. These findings suggest that multiple tautomeric states may be present in solution. To ensure a comprehensive analysis, we expand our study to include the alternative tautomeric form N2 of the cognate ligand by repeating the entire simulation workflow using this tautomer. Moreover, to obtain a complete perspective and enable comparison with previous findings, we carry out SMD simulations using both TIP4P-D and TIP3P water models, as well as WT-MetaD simulations,

for the cognate ligand in its N2 tautomeric form. The binding free energy computed for the N2 tautomer is notably lower than that of the previously studied N1 form. Importantly, these new results show much closer agreement with experimental data. In particular, simulations with the TIP4P-D water model yields a standard binding free energy of -13.2 ± 1 kcal/mol, and TIP3P yields -11.9 ± 1 kcal/mol. Both values align well with the experimental reference of -10.9 kcal/mol (Table 6.2). Additionally, these estimates are consistent with the outcome from the MetaD simulation in TIP4P-D, which reports a value of -13.9 ± 0.4 kcal/mol. A comprehensive summary of all binding free energy results for the cognate ligand is presented in Figure 6.21.

Interestingly, the N2 tautomeric state is less intuitive, as the cognate ligand in the N1 form structurally mimics the Guanine nucleobase, differing only by minor modifications. This close resemblance enables the N1 form to establish the canonical hydrogen-bonding network within the RNA binding pocket, particularly through Watson-Crick-like interactions with residue C15. Such behavior aligns well with Guanine’s known interaction pattern in this three-dimensional context (Figure 6.20B). In contrast, as illustrated in Figures 6.20B-C, the N2 form of the cognate ligand loses key hydrogen bonds. Specifically, two interactions with residues C15 and A29 are no longer present. Additionally, analysis of 500 ns plain MD simulations (two replicas) reveal distinct hydrogen bonding patterns for the two tautomers within the binding pocket over time. Figures 6.22 and 6.23 show these patterns for the two forms. Specifically, in the N1 tautomer simulations (Figure 6.22), the number of hydrogen bonds remains stable at approximately seven, whereas in the N2 tautomer (Figure 6.23), the average number of hydrogen bonds is five. These observations on the hydrogen bond network may help explain the lower binding affinity calculated for the N2 tautomeric form.

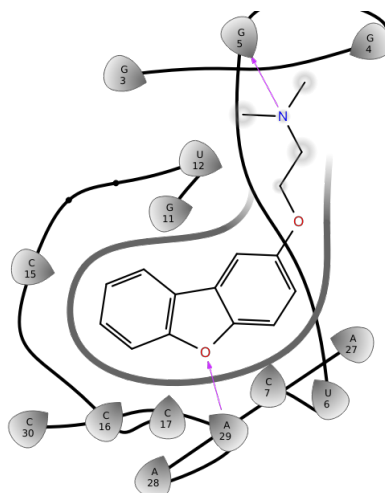


Figure 6.18: 2D representation of the hydrogen bonds established by the synthetic ligand within the riboswitch-preQ1 complex.

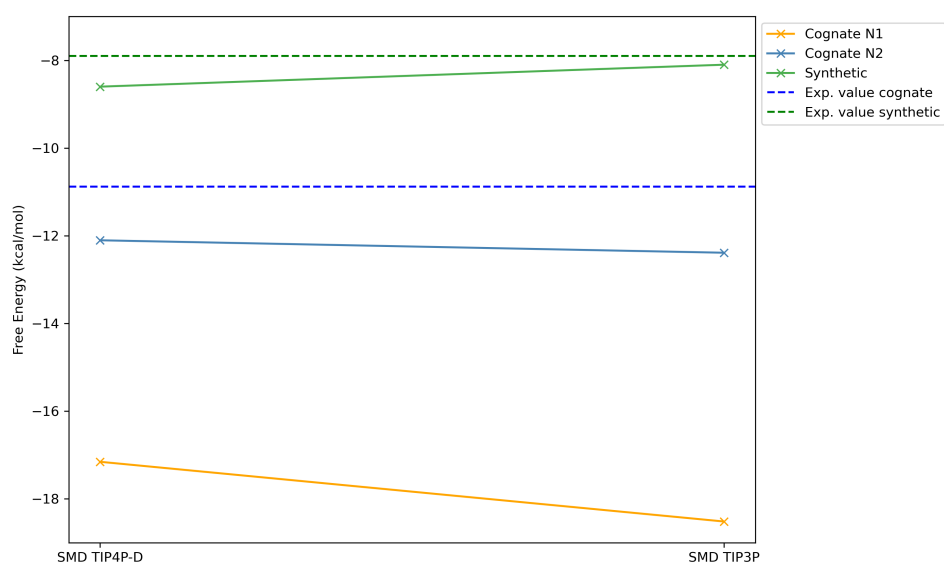


Figure 6.19: Binding free energies results. Dashed lines indicate experimental values (green: synthetic, blue: cognate). The x-axis represents the calculation method (SMD in TIP4P-D and TIP3P), while the y-axis shows the corresponding binding free energy. Ligands are color-coded: synthetic (green), cognate with AM1-BCC charges (orange) and cognate tautomer (light blue).

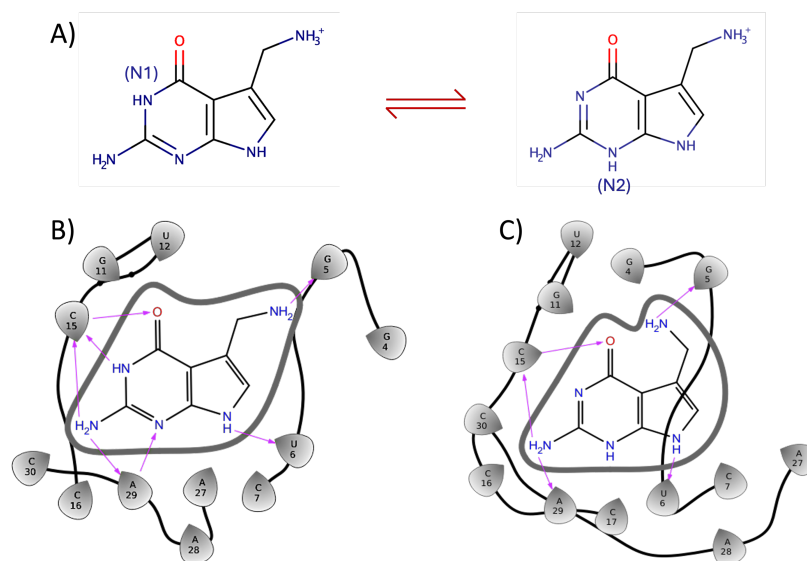


Figure 6.20: (A) Structures of the two tautomeric forms of the cognate ligand: N1 (left) and N2 (right). (B) Hydrogen bonding networks formed between the RNA target and the cognate ligand in the tautomeric form N1 and (C) in the tautomeric form N2. The two tautomeric forms establish a different hydrogen networking with the RNA target.

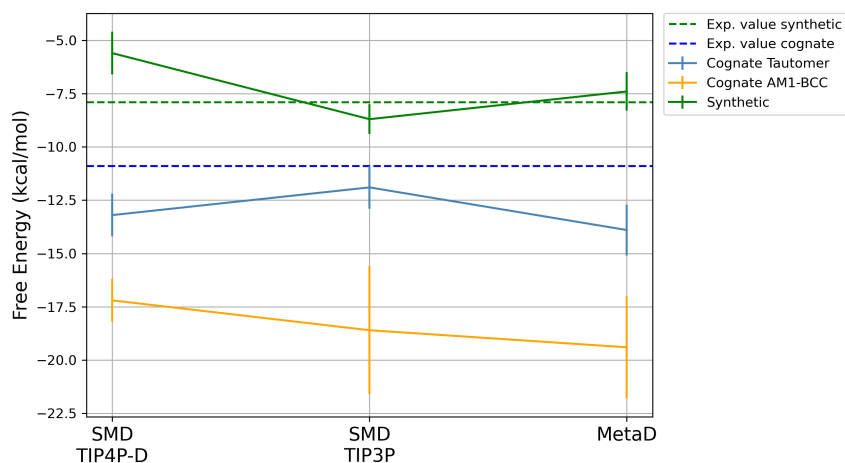


Figure 6.21: Binding free energies for all complexes tested. Dashed lines indicate experimental values (green: synthetic, blue: cognate). The x-axis represents the calculation method (SMD in TIP4P-D and TIP3P or MetaD), while the y-axis shows the corresponding binding free energy with associated errors. Ligands are color-coded: synthetic (green), cognate with AM1-BCC charges (orange) and cognate tautomer (light blue).

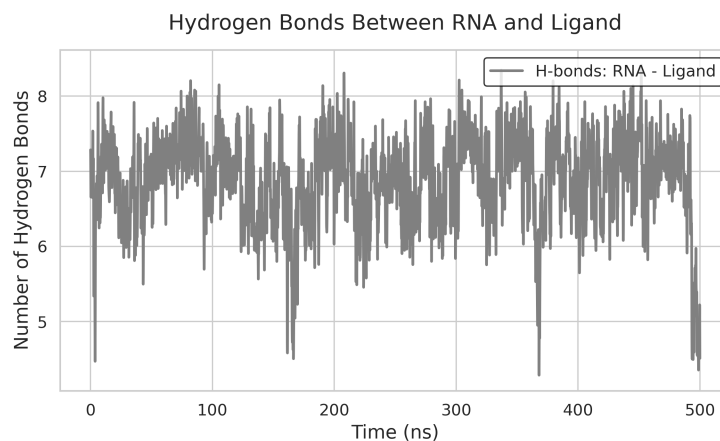


Figure 6.22: The number of hydrogen bonds for the N1 tautomer remained consistently around seven throughout the 500 ns of plain MD simulations.

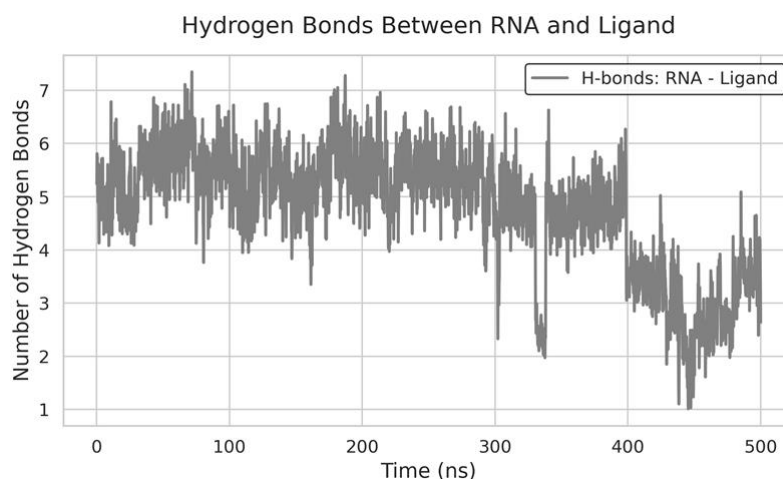


Figure 6.23: The number of hydrogen bonds for the N2 tautomer remained consistently around five throughout the 500 ns of plain MD simulations.

Ultimately, the coexistence of both tautomeric forms in solution may account for the observed discrepancy between the experimental binding affinity of the cognate ligand and the simulation results obtained for the N1 tautomer alone. Since both N1 and N2 are likely present in equilibrium in solution, the experimentally measured binding free energy likely reflects an average over both species. As a result, the presence of the lower-affinity N2 tautomer could reduce the overall observed affinity, leading to a value that is lower than what would be predicted by simulations considering only the more strongly binding N1 form.

6.5 Conclusions

In this applicative and conclusive study, we tackle the challenge of estimating binding free energies in complex and pharmaceutically relevant biomolecular systems using the nonequilibrium pipeline developed at the outset of my PhD work (Section 4 [P1]). Here, our investigation centers on two case studies: the binding of the anticancer drug Gleevec to its protein target Abl, and the interaction of two distinct ligands with the preQ1 RNA riboswitch. These systems exemplify typical challenges in computational drug discovery, ranging from the substantial size and conformational flexibility of the Abl-Gleevec complex to the dynamic nature and structural plasticity of the RNA receptor.

Building on our recent findings on minimizing dissipation during nonequilibrium simulations (Section 5 [P2]), we refine the strategy for constructing reference binding pathways used to define PCVs. A key improvement is the explicit inclusion of all molecular degrees of freedom essential to the binding process, specifically atoms from both the ligand and the receptor’s binding pocket. This enhancement proves critical for accurately capturing the structural rearrangements occurring in the binding sites. It results particularly relevant for the Abl-Gleevec complex, which involves a complex induced-fit mechanism, as well as for the riboswitch-ligand systems, where RNA flexibility plays a major role. With these improved reference pathways, we successfully applies our PCV-based protocol to estimate standard binding free energies across both protein- and RNA-ligand systems.

Moreover, consistent with the findings discussed in Section 5, our results reinforce the conclusion that the less dissipative TIP3P water model is better suited for nonequilibrium simulations. Its use led to binding free energy estimates that aligned more closely with experimental data, outperforming the more complex four-site TIP4P-D model in this context. Conversely, when applied within the framework of WT-MetaD, the four-site TIP4P-D water model yields results that are consistent with experimental observations. This contrast underscores the higher sensitivity of nonequilibrium SMD to the kinetic properties of the system, emphasizing the critical role of choosing an appropriate water model based on the specific

simulation methodology employed.

In conclusion, even-if equilibrium-based approaches remain the standard for free energy calculations in molecular simulations [256], nonequilibrium methods still pose methodological challenges but may offer distinct computational advantages. A key strength of our nonequilibrium pipeline is its inherent capacity for straightforward parallelization, which contrasts with more sequential techniques like MetaD. While the overall simulation time for the nonequilibrium SMD approach is longer than that of MetaD, all trajectories are independent and can be run in parallel. As a result, when utilizing high-performance computing infrastructure, the total time-to-solution is effectively reduced to the duration of a single trajectory, 100 ns for the systems examined in this final study. Moreover, our workflow enables a more direct and transparent evaluation of convergence, which is often less accessible in MetaD simulations. Finally, although the approach is computationally more intensive than widely used and efficient "end-point" methods (discussed in Section 2.5.1), it provides additional advantages. Beyond yielding reliable estimates of binding free energies, our protocol offers mechanistic insights into the ligand binding process and the identification of intermediate states along the pathway. These features are particularly valuable in pharmaceutical research, where understanding the dynamics and energetics of molecular recognition can significantly enhance rational drug design strategies.

Chapter 7

NanoShaperWeb: molecular surface and pockets detection made visual [P4]

7.1 Aim and Objectives

In the last three chapters, I presented our novel computational pipeline for calculating binding free energies in the nonequilibrium regime, detailing its advantages as well as its limitations. We conducted an in-depth analysis of the trade-off between computational efficiency and accuracy, and investigated the sources of dissipation that hinder convergence during nonequilibrium simulations. While we propose practical guidelines to minimize dissipation, we also challenged the methodology on highly complex protein systems and RNA complexes. These efforts contribute to establishing a robust nonequilibrium framework for binding free energy calculations, leveraging the inherent advantages of such approaches (such as parallelizability) while also yielding valuable mechanistic insights from the simulations.

Differently, this chapter presents a complementary project in the field of computational drug discovery, focusing on the development of a web-based interface for the NanoShaper code, aimed at molecular surface calculation and binding pocket

identification.

In structural biology and drug discovery, identifying protein cavities and analyzing molecular surfaces are crucial tasks that aid in functional site prediction and rational drug design. Over the years, various surface and pocket-detection tools have been developed, with NanoShaper being one of the available options. To enhance its accessibility, we introduce NanoShaperWeb, a web-based tool that integrates NanoShaper into an interactive user-friendly interface. This platform enables molecular surface and pocket detection, provides DrugPred descriptors [257] for deeper pocket analysis, and facilitates result visualization and download. By streamlining molecular analysis, NanoShaperWeb offers an efficient and user-friendly solution for researchers, supporting key stages of the drug development pipeline. Typical execution times are relatively short and depend primarily on the size and complexity of the system under analysis. For simple systems, such as the TRY-BEN complex, the analysis is usually completed within a few seconds. For larger or more complex biomolecular systems, execution times may increase accordingly. In addition, since NanoShaperWeb operates through a queue-based submission system, waiting times can occasionally extend the overall turnaround, depending on server load and user demand.

In the work presented thus far, NanoShaper has been employed for multiple purposes, including the evaluation of the standard volume correction term for $\Delta G_{binding}^{\circ}$ calculation [P1, P2, P3] and the examination of cryptic cavities in T4 lysozyme L99A/M102Q bound to N-phenylglycinonitrile [P3]. The NanoShaperWeb tool facilitates these analyses, thereby supporting our group’s ongoing and future research efforts.

The NanoShaperWeb tool is freely available online at <https://nanoshaperweb.iit.it/>, without the need for registration.

7.2 Introduction

As thoroughly explained in Chapter 1, the drug discovery process is complex and time-intensive [258], often relying on a multi-step pipeline that integrates various

computational methods [11]. Particularly in the early stages, computational modeling plays a crucial role in analyzing molecular surfaces and identifying cavities, aiding in the detection of pockets capable of accommodating small therapeutic molecules.

Proteins are among the primary molecular targets in drug discovery, as they play pivotal roles in numerous human cellular processes. Exogenous small molecules capable to bind proteins (i.e., ligands), are designed to bind specific protein pockets, thereby modulating their function and activity. Structure-based CADD leverages known protein structures to identify binding sites and support the design of new targeted therapeutics.

In this context, computational approaches for analyzing molecular surfaces and detecting protein cavities are essential for guiding drug discovery efforts. Among the many available pocket identification algorithms, typically classified as either geometric or energy-based, NanoShaper [201] emerges as a reliable tool for analyzing molecular surfaces and detecting protein cavities.

In this work, we introduce NanoShaperWeb, an application developed on the foundation of the original NanoShaper code. It is designed to offer an easy-to-use interface for accessing NanoShaper functionalities. Moreover, we integrate a NanoShaper-based implementation of the DrugPred descriptors [257], offering a rapid computation of both geometric and chemical descriptors for each identified pocket. This provides a comprehensive understanding of the pockets’ characteristics and may serve as foundation for machine learning approaches.

7.3 Implementation

We have modified the original NanoShaper source code, transforming it into a more flexible and structured software. Additionally, we have incorporated the descriptor builder [259] for generating DrugPred descriptors [257] within the code. The NanoShaperWeb application is designed using a microservices architecture consisting of three key components: the NanoShaper executable, a Flask-based web server, and the Browser App. Flask manages back-end processing of requests and

response handling, while the Browser App offers an interactive interface. The three components work together to create a fully integrated web application. Moreover, all components are containerized with Docker, ensuring reliable deployment and scalability. An overview of the system architecture is shown in Figure 7.1.

7.3.1 NanoShaper and Descriptors

The NanoShaperWeb application is built around the two core algorithms of the original NanoShaper code: surface computation and pocket detection. It also offers visualization for all results generated by NanoShaper, along with the additional DrugPred descriptors.

Surfaces: NanoShaper provides the capability to triangulate various molecular surface definitions [260, 261]. In general, NanoShaper operates by embedding the protein into a cubic grid and triangulating the molecular surface using ray-casting. Rays are cast along grid edges to analytically sample the surface while simultaneously obtaining the in/out information required for triangulation. All the molecular surfaces are generated through the NanoShaper interface by running the tool multiple times. NanoShaper supports the following surfaces: van der Waals surface (WDW), skin surface [262], blobby surface [263], and Connolly-Richards surface or solvent-excluded surface (SES).

Pocket detection: Within the algorithm, the region of space that is accessible to a water molecule but not to a larger spherical probe is defined as a pocket. This pocket is identified by calculating the volumetric difference between two regions defined by two distinct solvent-excluded surfaces. These SESs are computed using different probe radii, namely R and r where $R > r$. This volumetric difference is calculated over the grids, followed by a clean-up step to eliminate noisy grid points, and then a flood-filling algorithm is applied to isolate the pockets. The definition of the pocket shapes is affected by the size of the probe: a smaller radius r smooths gaps in the inner surface, while a bigger radius R helps to identify shallow pockets. The default values used are $R = 3.0 \text{ \AA}$ and $r = 1.4 \text{ \AA}$. The smaller radius $r = 1.4 \text{ \AA}$ is selected to represent the approximate size of a molecule of water. Additionally, NanoShaper is capable of detecting closed cavities within

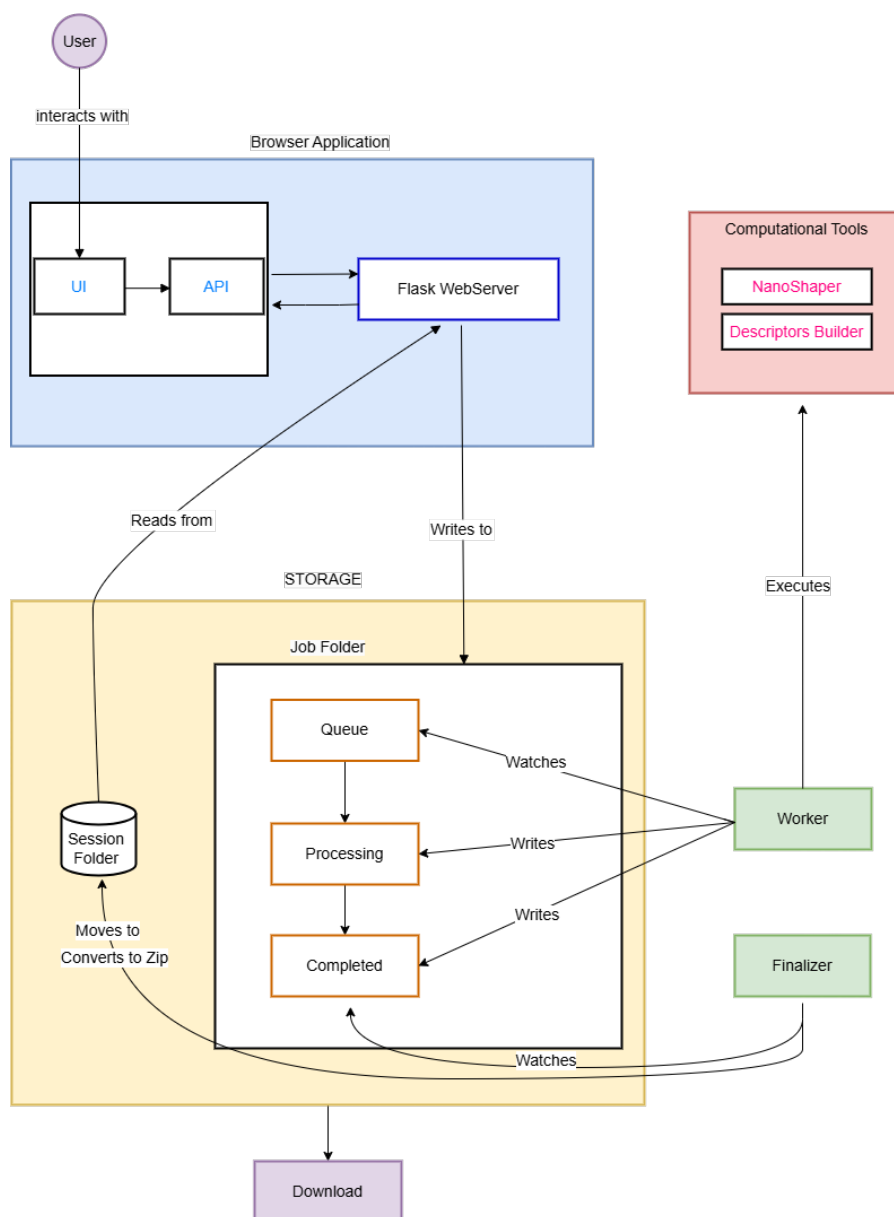


Figure 7.1: The chart depicts the architecture of NanoShaperWeb, highlighting the interactions between the user, browser application, storage system, computational tools, and worker processes. Users interact with the web interface, which communicates with a Flask-based server. Submitted tasks are stored in a job folder and processed via a queue system. Worker processes handle computational tasks, utilizing tools like NanoShaper and Descriptors Builder. Once the computations are complete, results are saved and made available for users to download.

the surface and estimating their volume. In NanoShaper, pockets and cavities are distinct objects, namely pockets have a defined body and entrance, while cavities do not exhibit such structure.

Descriptors builder: subsequently, every pocket is characterized by the entrance area and points provided by NanoShaper, as well as by geometric and chemical descriptors [257]. The descriptor computation is done through a Python script running on top of the NanoShaper executable. The complete list of calculated descriptors and their corresponding abbreviations is presented in Table 7.1. In this set of descriptors, those related to amino acid composition are determined based on the frequency of different amino acid classes, categorized by their physicochemical properties. The occurrence of each amino acid type (in_X) is represented as a descriptor, calculated by summing the surface areas surrounding the amino acid X . Moreover, pocket properties include surface areas for hydrogen-bond donors and acceptors (dsa_t and asa_t), which are calculated around polar atoms. By subtracting the donor and acceptor areas from the total surface area, the hydrophobic surface area (hsa_t) is also derived. Additionally, to provide relative measurements, descriptors (dsa_r , asa_r , and hsa_r) are obtained by normalizing each value by the total binding site surface area. The relative polar surface area (psa_r) is defined as the sum of dsa_r and asa_r . Lastly, a compactness descriptor is included to characterize cavity and pocket shapes.

7.3.2 Web Server

The NanoShaperWeb server is developed using Flask, a lightweight Python web framework to manage HTTP requests, user sessions, PDB file processing, and job execution. For molecular structure manipulation and preprocessing, the server uses the ProDy library [264, 265]. This includes parsing PDB files with `parsePDB` for structure loading, applying atom subselections using VMD-like syntax, eliminating duplicate atoms, and converting structures into the .xyzr format required by NanoShaper. To maintain consistency in atom naming, Amber FF parameter files are used.

To optimize performance for computationally demanding tasks, some routines are

Descriptor	Name
Binding site volume	<i>vol</i>
Total surface area	<i>area_b</i>
Entrance area	<i>area_e</i>
Binding site compactness	<i>cness</i>
Relative hydrogen-bond donor surface area	<i>dsa_r</i>
Hydrogen-bond donor surface area	<i>dsa_t</i>
Relative hydrogen-bond acceptor surface area	<i>asa_r</i>
Hydrogen-bond acceptor surface area	<i>asa_t</i>
Relative hydrophobic surface area	<i>hsa_r</i>
Hydrophobic surface area	<i>hsa_t</i>
Relative occurrence of polar amino acids	<i>paa</i>
Relative occurrence of non-polar amino acids	<i>haa</i>
Relative occurrence of multifunctional amino acids	<i>maa</i>
Relative occurrence of charged amino acids	<i>caa</i>
Relative polar surface area ($dsa_r + asa_r$)	<i>psa_r</i>
Incidence of amino acid X in the binding site relative to the surface	<i>in_X</i>

Table 7.1: Descriptors and their corresponding names.

implemented in C and enhanced with AVX2 vector instructions when available. The backend architecture follows a modular design with five key components, including the API Layer, Session Management, Job Queue System, Worker Process, and Finalizer Process. Further technical details can be found in our application note [P4].

7.3.3 Browser App

The browser app serves as the client-side interface, enabling user interaction with the web server through asynchronous HTTP requests. It is developed using modern web technologies and follows a structured design for efficient usability and performance. The implemented frameworks and components are following summarized.

Frontend Framework: uses HTML5, TailwindCSS for styling, and vanilla JavaScript to handle interactivity. The layout is organized into distinct sections for configuring inputs, managing jobs, and displaying results.

3D Visualization: the molecular visualization (incorporating custom wrappers)

is obtained by employing 3DMol.js [266]. It supports interactive selection, multiple rendering modes, and different pocket coloring.

State Management: stores visualization settings, selected pockets, and UI preferences on the server filesystem, synchronizing client-side updates in real time. Session data, including viewer states, is saved in JSON files on the server, ensuring users can resume analysis across multiple sessions.

Web Workers Architecture: uses dynamically generated web workers for multi-threaded client-side processing, handling resource-intensive tasks like large file downloads while tracking progress. This approach maintains UI responsiveness and includes fallback mechanisms if web workers are unavailable.

The application features a responsive design, ensuring compatibility across different devices and screen sizes. Results can be downloaded as ZIP directories containing explanatory log files, surface meshes, pocket and cavity PDB files, and descriptor data in convenient CSV format for successive analysis. Finally, the entire application is containerized using Docker.

7.4 Web server interface

The interface of NanoShaperWeb front page is shown in Figure 7.2. As can be seen, the bottom panel of the web server provides three links to individual pages: *Find Pockets!*, *About NanoShaper* and *Pocket Descriptors*.

The *About NanoShaper* page offers a brief overview of the underlying methods, while the *Pocket Descriptors* page provides a summary of the DrugPred descriptors [257] computed using NanoShaper. In the *Find Pockets* page, users can either initiate a new session or resume a previous analysis by entering a session ID. When starting a new session, users can submit a job by downloading from the web or uploading a PDB/PDBx-mmCIF file.

Additionally, key information is displayed through tooltips, allowing users to access helpful information by hovering over specific elements in the NanoShaperWeb interface.

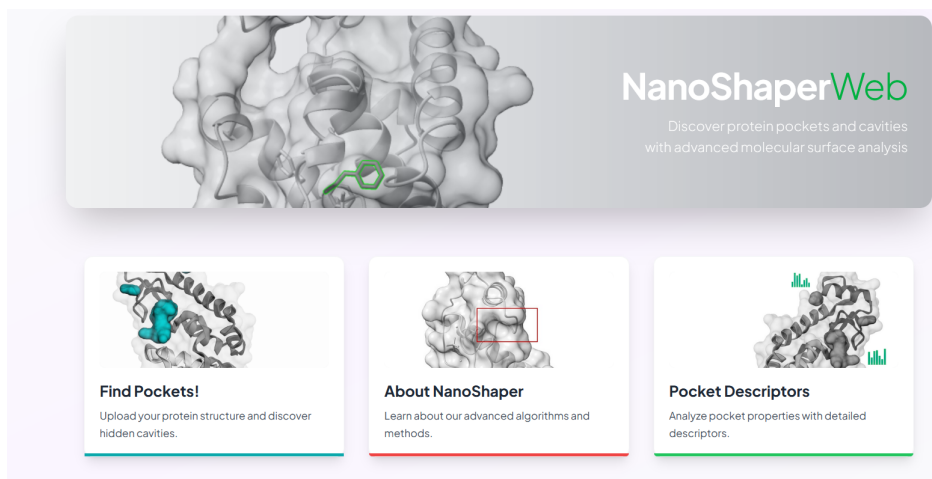
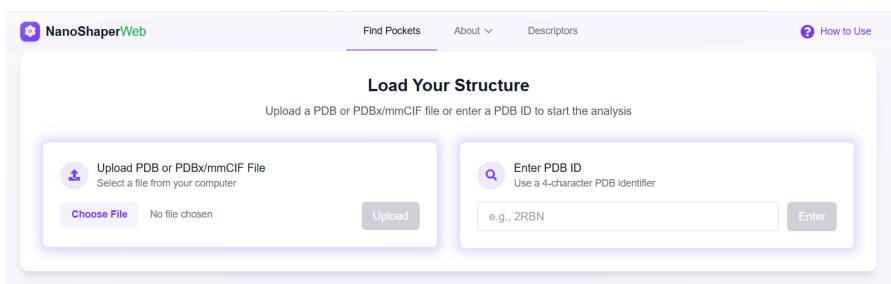


Figure 7.2: Interface of the NanoShaperWeb front page

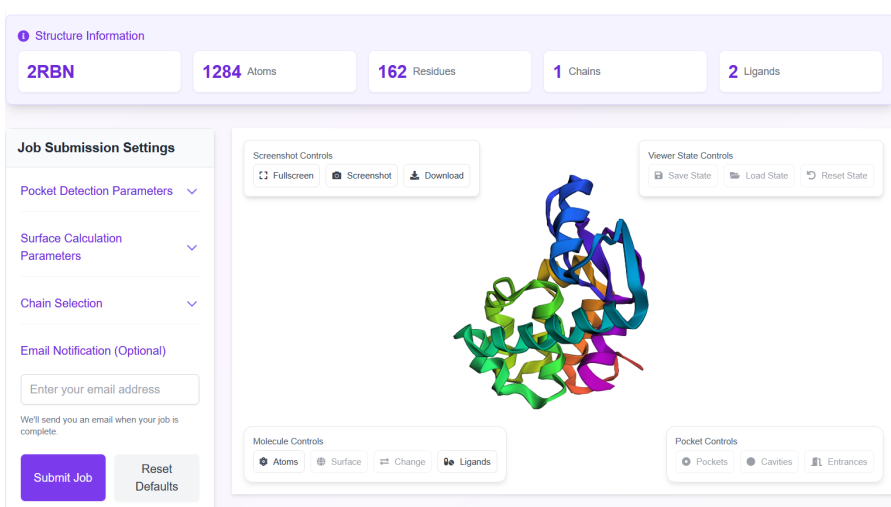
7.4.1 Input, Settings, Viewer and Pocket Table

The usage of NanoShaperWeb is rather intuitive, and it is illustrated in the *How to Use* section within the web application. As a representative example to demonstrate the usage of NanoShaperWeb, we present the structure of T4 lysozyme L99A/M102Q bound to N-phenylglycinonitrile (PDB code: 2RBN) previously analyzed in Chapter 5 [P2]. Specifically, Figure 7.3 represents the inputs, while Figure 7.4 reports the output of this system.

As anticipated, in the *Find Pockets* page, users can retrieve a previous analysis (using the session ID) or start a novel analysis with a new session. Selecting the latter option opens the input section, where the main input (uploaded or downloaded) is the protein file. The PDB/PDBx-mmCIF protein file undergoes preprocessing, and is automatically displayed in molecular cartoon mode (Figure 7.3b). Interesting structural details such as the number of residues, chains, and atoms are also displayed. Moreover, users can customize the molecular visualization by interactively change representation, highlighting ligands or selecting specific chains to better suit their needs. By simply clicking *Submit*, the standard analysis can start and it appears in the *Jobs Management* feed. Here, users can verify the analysis status, the queue position and execute actions such as view or download results. Additionally, in the *Job Submission Settings* tab, advanced users can run analysis



(a) In the *Find Pockets* page, users can start a new job session. Here, users can specify the protein for analysis by uploading or downloading the PDB of interest. As a representative example, the 2RBN protein is selected.

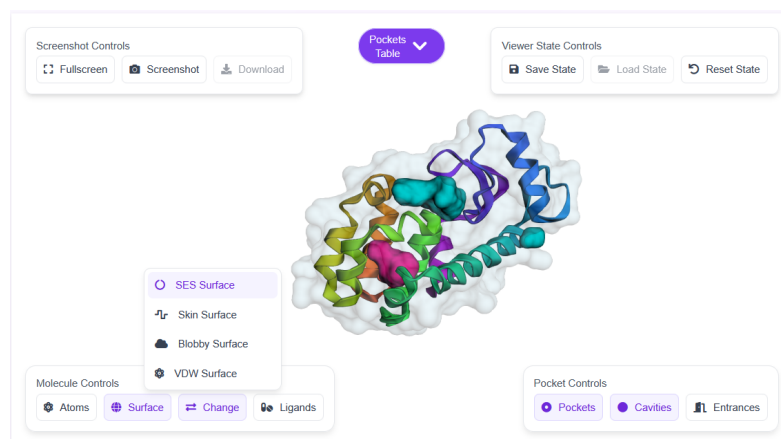


(b) The PDB file has been successfully downloaded and is ready for analysis. Key structural information is displayed, and users can interactively adjust the PDB visualization (e.g., atom-based views). Clicking the "Submit Job" button initiates the pocket detection process, placing the job in the queue with a unique job ID. The "Job Submission Settings" tab allows users to fine-tune parameters for more precise control over the analysis.

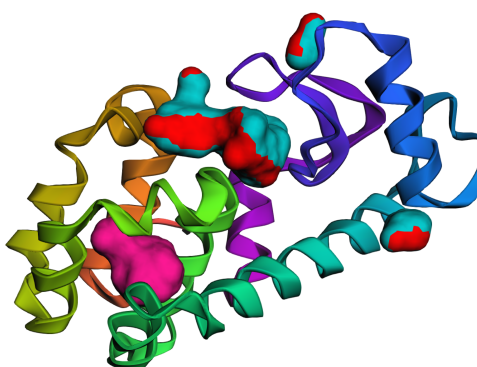
Figure 7.3: User interface of the *Find Pockets* page, with a visual walkthrough of how to initiate a new analysis session.

only on a subset of the chains, or they can set custom parameters such as the SES probe radii and the Skin smoothness. In the *Jobs Management* feed, once the submitted job is finished, users can visualize the results by clicking the *View* button. This action loads the results into the interactive viewer and the pockets table components.

At the heart of the application is the molecular viewer, providing an interactive display of molecular structures as shown in Figure 7.4a. The viewer offers intuitive



(a) After the analysis, the protein can be visualized using four different surface representations: Connolly-Richards surface, skin surface, blobby surface, and van der Waals (VdW) surface. Pockets and cavities are highlighted by default in distinct colors for easy identification.



(b) Fullscreen screenshot of the 2RBN protein visualization.

Figure 7.4: Visualization of NanoShaper output for the 2RBN protein, highlighting different surface types and pocket detection features.

3D interaction using standard mouse controls, allowing users to easily rotate, pan, and zoom the molecular structure. Surrounding it there are several control panels, each serving a dedicated purpose. An opacity slider allows users to adjust surface transparency in real time, while the *Molecule Controls* panel offers options to toggle molecular surfaces and visualize individual atoms. In the bottom-right corner, the *Pocket Controls* panel provides options to toggle the visibility of pockets and cavities. In addition, a dedicated visualization mode highlights pocket entrances in red, distinguishing them clearly from the rest of the pocket structure, as shown in Figure 7.4b. This distinction between the pocket entrance and body is particularly valuable, as it helps identify potential access points to binding sites. All visibility settings, color configurations, and selected pockets used during the analysis can be saved using the tools in the *Viewer State Controls* located in the top-right corner. This feature allows users to easily restore the entire viewer state at any time, ensuring consistent and reproducible analysis across different sessions. In the top-left corner, the *Screenshot Controls* panel allows users to capture high-resolution images of the current view. For a clearer and more detailed snapshot, the visualization area can be maximized using the fullscreen option before taking the screenshot.

Positioned below the viewer and designed to work in synergy with it, the *pockets table* serves as an interactive summary for pocket analysis and visualization. Users can add or remove individual pockets from the viewer by clicking the corresponding buttons in the table, and customize their appearance using the built-in color picker. Once the NanoShaper analysis is complete, the viewer initially loads only the 10 largest pockets or cavities by default. Researchers can then use the interactive pocket table to take full control over which pockets are displayed and how they appear. Each entry in the table presents key metrics such as surface area, type, and volume. Additionally, pockets and cavities can be sorted by any of these metrics simply by clicking on the corresponding column header. The pockets table allows for efficient multi-selection using Shift+click to select consecutive pockets and Ctrl+click to choose non-consecutive ones. Any modifications or selections made in the pocket table are instantly reflected in the viewer. This seamless inte-

gration between the tabular data and 3D visualization fosters an efficient workflow, enabling researchers to quickly explore and analyze potential binding sites within the molecular structure. For a more detailed analysis, each table entry features a *Show Details* button that reveals the full set of descriptors calculated by the server. These can easily be copied and pasted or downloaded directly from the outcomes.

7.4.2 Mail Notification and Results Download

In the *Job Submission Settings* tab, users can optionally enter an email address to receive automatic notifications upon job completion. These emails include important details, such as the session ID. Additionally, two links are provided: one to resume the session for immediate visualization and analysis, and another for direct downloading of the results. This email notification service ensures that researchers stay informed about the progress of their analyses, particularly for lengthy computations. Finally, all results are packaged into a well-organized ZIP directory, which can be downloaded either through the link provided in the notification email or directly from the webserver. The files within the ZIP archive comprehensively document the analysis and include Molecular Surface Files, Individual Pocket/Cavity PDB Files, Log Files, and Descriptor Data. Specifically, the Molecular Surface Files are provided for all surface types and include mesh representations of the computed surfaces, available in either MSMS or .off format. Individual PDB files are created for each detected pocket and cavity, while the log file serves as a comprehensive summary containing key information from the analysis (such as volume measurements, surface area calculations, and detailed lists of the constituent atoms for all internal cavities and pockets identified within the molecular system). The descriptor data is organized in a CSV file, presenting the descriptors for each pocket in a machine learning-ready format. This organization of the results ensures that users have access to both visual representations and the underlying quantitative data, providing everything needed for a thorough analysis of molecular surfaces and binding sites.

7.5 Conclusion

NanoShaperWeb improves access to the NanoShaper framework by offering a web-based platform that eliminates the need for installation. Its intuitive interface enables researchers to perform molecular surface analysis and pocket detection effortlessly, with computations executed remotely in parallel. The inclusion of DrugPred descriptors enhances pocket characterization, providing valuable insights into their properties. By simplifying molecular analysis, NanoShaperWeb serves as a powerful and accessible tool, aiding key stages of both my PhD research and of the drug design process.

Chapter 8

Absolute residence times estimation from scaled molecular dynamics simulations: application to Mcl-1 inhibitors

8.1 Aim and Objectives

This chapter describes a complementary project focused on ligand-binding kinetics, with a particular emphasis on the feasibility of accurately predicting not only relative but also absolute residence times. As described in Section 1.2.1, RT is a key parameter for the success of drug development. Here, we aim to introduce and validate a MD-based approach capable of both ranking compounds by their RTs and, importantly, estimating their absolute RT values. The ultimate goal is to validate this methodology against experimental data for a biologically relevant pharmacological system, Mcl-1 and its inhibitors, while striving for an optimal balance between computational efficiency and predictive accuracy. The corresponding article describing this work is currently under submission. Some results and graphical analyses are still under investigation and refinement; for this reason, they are not fully presented here and will be reported in detail in the final published work.

8.2 Introduction

In recent years, the crucial role of biomolecular kinetic parameters on the biological activity of small molecules has become progressively recognized. While binding affinity remains a key factor, drug efficacy is increasingly understood to depend on the RT which can influence both potency and therapeutic outcomes. As discussed in Chapter 1.2.1, the prediction and optimization of RT play a pivotal role in drug development, particularly during the early discovery phases [27, 13]. Therefore, the ability to computationally predict and optimize RT can significantly accelerate drug discovery. Particularly, structure-kinetics relationship studies have gained considerable attention, although the challenges in this domain remain larger than those in equilibrium free energy calculations [43].

While MD simulations are ideally suited for providing the atomistic detail needed to study binding kinetics, the long timescales involved make brute-force simulations impractical for routine RT estimation. This timescale problem, introduced in Section 2.4, has motivated the development of various enhanced sampling methods. Techniques such as MetaD, SMD, and US, which are reviewed in Section 2.4, have been applied to this challenge [27, 35, 43].

However, calculating absolute kinetic rates remains computationally expensive, often requiring microsecond-scale simulations for a single ligand, which is unfeasible for high-throughput screening. This has led to the development of simplified approaches for estimating relative RTs, which are valuable for ranking compounds. One valuable technique is scaled-MD, which accelerates unbinding by applying a smoothing parameter (λ) to the system’s potential energy surface, effectively ”flattening” the energy barriers. The degree of smoothing is controlled by λ , where lower values correspond to more aggressive scaling and faster unbinding. Previous work has demonstrated the utility of scaled-MD for ranking molecules based on their dissociation rates [267, 268].

In this study, conducted in collaboration with Servier, we explore the feasibility of extending the scaled-MD methodology to predict not only relative but also absolute RTs for a large set of 71 small-molecule inhibitors of myeloid cell leukaemia-1

(Mcl-1). Mcl-1 is a well-validated anti-apoptotic protein and an interesting therapeutic target in oncology. By applying our refined methodology to this extensive and pharmacologically relevant dataset, we aim to demonstrate a computationally efficient path towards absolute RT prediction.

8.3 Methods

The experimental data and initial scaled-MD simulation data for the 71 Mcl-1 inhibitors were previously generated through a collaboration with Servier. My contribution to this project focuses exclusively on the development and implementation of the post-processing algorithms for predicting relative and absolute residence times from the existing simulation data. The preparatory steps, including the use of in-house X-ray crystal structures and the specific scaled-MD protocol, are described in detail in the manuscript currently under submission.

Here, we develop two distinct methods for estimating the physical unbinding time from scaled-MD simulations. First, we introduce a simple, closed-form approximate formula for a quick estimation of absolute RTs. Second, we present a more rigorous and comprehensive algorithm that leverages simulations performed at multiple scaling values λ to deliver more robust estimates of absolute RTs. The statistical significance of all predictions is assessed using bootstrap analysis.

8.3.1 From Scaled to Relative Unbinding Times

As previously established [267], a scaled-MD simulation with a scaling factor λ is analogous, from a Boltzmann distribution perspective, to running the simulation at a scaled temperature $\tilde{T} = T/\lambda$. To recover relative physical RTs from the computed scaled unbinding times ($\tilde{t}_\lambda$), we assume an Arrhenius-like relationship for dissociation. By normalizing the scaled and experimental unbinding times with respect to a reference ligand (here, the fastest one), the relative physical RTs can be "unscaled" using the following relation:

$$\frac{\tilde{t}_{2,\lambda}}{\tilde{t}_{1,\lambda}} = \left(\frac{t_2}{t_1}\right)^\lambda \quad (8.1)$$

where the subscript 1 refers to the reference ligand and 2 refers to the ligand for which the relative RT is being determined.

8.3.2 From Scaled to Absolute Unbinding Times

Predicting absolute RTs requires a more robust theoretical framework to properly account for the biasing introduced by the potential scaling. Here, we use transition state theory and the Eyring–Polanyi equation introduced in Chapter 1 (Eq. 1.22). The physical residence time t (the inverse of k_{off}) is given by:

$$t = \frac{1}{k_{off}} = \frac{h}{k_B T} \exp(\beta \Delta G^\ddagger) \quad (8.2)$$

where $\Delta G^\ddagger$ is the Gibbs free energy of activation for the dissociation process. Our goal is to derive a clear relationship between the computationally observed scaled RT, $\tilde{t}$, and the physical RT, t .

8.3.3 Relationship between scaled and physical RT

Here, we derive a relationship between the physical time and the scaled time by analyzing the system’s potential within the harmonic approximation. Relying on the harmonic approximation, the harmonic potential of the unperturbed system is:

$$V(x) = V_0(x) + \frac{1}{2}(x - x_0)^t H(x - x_0) \quad (8.3)$$

where H is the Hessian matrix. The corresponding harmonic free energy is:

$$G = V_0 - k_B T \sum_{i=1}^{3N} \log \left(\frac{k_B T}{\hbar \omega_i} \right) \quad (8.4)$$

where ω_i are the harmonic modes of the system. For the scaled potential system, $\lambda V(x)$, the scaled angular frequencies become $\omega_i^s = \sqrt{\lambda} \omega_i$. The scaled free energy $\tilde{G}$ can thus be related to the unscaled free energy G :

$$\tilde{G} = G + V_0(\lambda - 1) + \frac{3}{2} N \log \lambda \quad (8.5)$$

From this, we can relate the scaled and unscaled free energy barriers of activation, $\Delta\tilde{G}^\ddagger$ and $\Delta G^\ddagger$:

$$\Delta\tilde{G}^\ddagger = \Delta G^\ddagger + \Delta V(\lambda - 1) \quad (8.6)$$

By substituting this expression into the Eyring-Polanyi equation for the scaled system, the scaled residence time $\tilde{t}$ becomes:

$$\tilde{t} = \frac{h}{k_B T} \exp\left(-\beta \Delta\tilde{G}^\ddagger\right) \quad (8.7)$$

$$= \frac{h}{k_B T} \exp\left(-\beta(\Delta G^\ddagger + \Delta V(\lambda - 1))\right) \quad (8.8)$$

Considering Eq. 8.2 for the physical unbinding time t , we arrive at a direct relationship between the scaled and physical RT:

$$\tilde{t} = t \exp(\beta \Delta V(\lambda - 1)) \quad (8.9)$$

This equation relates the scaled unbinding time $\tilde{t}$, recovered in scaled MD simulation, to the physical absolute unbinding time t . It forms the theoretical basis for our absolute RT estimators.

Simple Enthalpic Approximation

As a first approach, we propose a simplified, single- λ enthalpic approximation. We assume that the free energy barrier scales linearly with the potential, $\Delta\tilde{G}^\ddagger \approx \lambda \Delta G^\ddagger$, and that the simulation at a scaled potential is equivalent to a simulation at a scaled temperature $\tilde{T} = T/\lambda$. This leads to the expression for the scaled residence time:

$$\tilde{t} \approx \frac{h}{k_B \tilde{T}} \exp(\beta \Delta\tilde{G}^\ddagger) = \lambda \frac{h}{k_B T} \exp[\beta \lambda \Delta G^\ddagger] \quad (8.10)$$

By rearranging this equation and using the definition of the physical time t (Eq. 8.2), we can solve for t :

$$t = \frac{\tilde{t}^{1/\lambda}}{\lambda^{1/\lambda} \left(\frac{h}{k_B T}\right)^{\frac{1-\lambda}{\lambda}}} \quad (8.11)$$

This provides a straightforward, closed-form equation to estimate the absolute physical RT from a single set of scaled-MD simulations performed at a given λ .

Multiple Scaling Method

To obtain a more robust estimate, we developed a method that leverages multiple scaled-MD simulations performed at different λ values. This approach is analogous to how the Arrhenius constant is determined experimentally from reactions run at different temperatures.

Starting from a slightly different derivation based on the harmonic approximation of free energy, we obtain the following relationship between scaled and physical RT:

$$\tilde{t} = \lambda t \exp[\beta \Delta V^\ddagger (\lambda - 1)] \quad (8.12)$$

where $\Delta V^\ddagger$ is the potential energy of activation. While λ is a known simulation parameter, $\Delta V^\ddagger$ is unknown. To solve for both t and $\Delta V^\ddagger$, we can perform simulations at multiple λ values.

Taking the natural logarithm of Eq. 8.12, we get a linear equation for each ligand l and each scaling factor λ_i :

$$\ln(\tilde{t}_{l,i}) = \ln(\lambda_i) + \ln(t_l) + \beta \Delta V_l^\ddagger (\lambda_i - 1) \quad (8.13)$$

This can be rewritten as $\gamma_{l,i} = \nu_l + \beta \Delta V_l^\ddagger (\lambda_i - 1)$, where $\gamma_{l,i} = \ln(\tilde{t}_{l,i}/\lambda_i)$ is the known quantity from simulations and $\nu_l = \ln(t_l)$ and $\Delta V_l^\ddagger$ are the two unknowns for each ligand.

For a set of N_L ligands simulated at n_λ different scaling factors, we can construct an over-determined system of linear equations. We can formulate this using absolute relationships (Eq. 8.13) for each simulation, as well as relative relationships between pairs of ligands. The full system of equations can be expressed in matrix form:

$$\bar{\gamma} = \mathbf{A} \bar{x} \quad (8.14)$$

where $\bar{\gamma}$ is the vector of known simulation results, $\mathbf{A}$ is the design matrix con-

structed from the known λ values, and $\bar{x}$ is the solution vector containing the unknown $\ln(t_i)$ and $\Delta V_i^\ddagger$ values for all ligands.

Given that this system is ill-conditioned, we solve it using Tikhonov regularization to ensure a numerically stable solution:

$$\min_x ||\mathbf{A}\bar{x} - \bar{\gamma}||^2 + \theta ||\bar{x}||^2 \quad (8.15)$$

The closed-form solution is:

$$\bar{x} = (\mathbf{A}^T \mathbf{A} + \theta \mathbf{I})^{-1} \mathbf{A}^T \bar{\gamma} \quad (8.16)$$

The optimal regularization parameter θ is determined automatically using the generalized cross-validation (GCV) method. This approach allows for the simultaneous and robust determination of the absolute physical RTs for all ligands in the subset for which multiple- λ data is available.

The complete numerical derivations, along with additional considerations and analyses, are provided in the manuscript currently under submission.

8.4 Results and Discussion

We applied our computational methods to the dataset of 71 Mcl-1 inhibitors, which span a wide range of experimental RTs from 2 to 523 seconds.

8.4.1 Calibration of the Scaling Factor

For the single- λ analysis of the full 71-compound set, an optimal scaling factor was required. Four λ values (0.36, 0.40, 0.42, 0.45) on a subset of 7 diverse compounds were tested. Values below 0.45 resulted in a diffusive regime with poor discrimination, while higher values led to computationally prohibitive simulation times. Therefore, $\lambda = 0.45$ was chosen as the optimal balance between computational cost and accurate kinetic ranking for the entire series. Twenty independent scaled-MD simulations were then run for each of the 71 complexes using this scaling factor.

8.4.2 Relative and Absolute Unbinding Time Predictions

First, we compute the scaled RTs ($\tilde{t}$) and we confirm a strong correlation with experimental data (Pearson’s $R = 0.85$, $R^2 = 0.72$). We then calculate relative RTs using the Arrhenius-like correction (Eq. 8.1), which maintains the same excellent correlation, validating the approach for ranking compounds.

Next, we apply the novel absolute RT estimators.

Simple Enthalpic Approximation: Using Eq. 8.11, we convert the scaled RTs from the $\lambda = 0.45$ simulations into absolute physical RTs. This simple method preserves the high correlation with experimental data (Pearson’s $R = 0.85$, $R^2 = 0.72$) and yields absolute RTs in the correct order of magnitude. The mean relative error across the 71 compounds is 0.8, demonstrating a good predictive capability for such a straightforward and computationally inexpensive method.

Multiple Scaling Method: We apply the more rigorous multiple scaling method to the subset of 7 compounds for which we have simulation data at four different λ values. We solve the regularized system of equations (Eq. 8.16) to simultaneously find the absolute RTs. This approach yields good correlations with experimental data, with Pearson’s R values as high as 0.86. Moreover, when we treat one compound, known to be in a non-optimal docked pose, as an outlier, the correlation for the remaining 6 compounds improves dramatically to $R = 0.97$. This highlights the high accuracy of the method when provided with high-quality input structures and multi- λ data.

A comparison of the different methods (Table 8.1) shows that while the simple enthalpic approximation provides a good balance of speed and accuracy for large-scale screening, the multiple scaling method offers a path to even higher accuracy for smaller, focused sets of compounds, confirming its potential for lead optimization stages. A more detailed conclusive table, together with an in-depth analysis, is provided in the manuscript.

Finally, the predicted RTs have been used to analyze the structure-kinetics relationships within the chemical series. By projecting the compounds onto similarity

Table 8.1: Summary of correlation results for the benchmark subset (7 compounds). "No outlier" refers to the calculation on 6 compounds, excluding a known non-optimal pose.

Method	R (log vs log)	R (no outlier)
Simple Enthalpic Approx.	0.72	0.95
Multi-scaling (Abs+Rel)	0.86	0.97
Multi-scaling (Abs+Rel) + Prior	0.86	0.97

maps, we confirm that our predictions correctly cluster compounds with high and low RTs, mirroring the experimental data and identifying "kinetic cliffs"-pairs of structurally similar compounds with different kinetics.

8.5 Conclusion

In this study, we successfully develop and validate a computational methodology based on scaled-MD simulations for the prediction of ligand residence times. We demonstrate that:

1. A simple, Arrhenius-like correction can be used to reliably rank a large series of congeneric ligands, achieving excellent correlation with experimental data.
2. We introduce a novel, closed-form "Simple Enthalpic Approximation" that allows for the straightforward and rapid estimation of absolute residence times with a remarkable degree of accuracy, making it suitable for large-scale compound screening.
3. We develop a more rigorous "Multiple Scaling Method" that leverages data from simulations at several scaling factors to deliver highly accurate absolute RT predictions, ideal for in-depth analysis during lead optimization.

By applying these methods to a challenging and extensive dataset of 71 Mcl-1 inhibitors, we shown that computationally efficient approaches can move beyond relative ranking to provide quantitative, absolute predictions of a crucial kinetic parameter. This work represents a significant step towards the integration of predictive kinetics into the early stages of the drug discovery pipeline.

Chapter 9

Supporting Information

This chapter collects supplementary results, analyses, and methodological details that support the main body of the Thesis. In order to avoid overloading individual chapters with extensive data and technical information, these additional materials have been gathered here. The information presented in this chapter is intended to provide the necessary depth to fully understand and reproduce each step of the work, while preserving the readability of the main chapters. Where relevant, references to this Supporting Information are made throughout the Thesis to guide the reader to detailed results and complementary discussions.

9.1 Minimizing dissipation is nonequilibrium binding free energy simulations

9.1.1 Well-Tempered MetaDynamics

To ensure comprehensive validation of our findings, we also perform WT-MetaD simulations. This method provides an interesting point of comparison, as it remains a nonequilibrium technique due to its use of a time-dependent Hamiltonian. However, unlike approaches that rely on CFT or JE estimators, it does not require the calculation of work values.

MODEL	TIP3P	OPC3	TIP4P	OPC
Height	0.2	0.5	0.5	0.5
Bias Factor	10	15	15	15
Peace	750	500	500	500
Binding Free Energy	-4.108	-6.966	-6.853	-6.691

Table 9.1: WT-MetaD parameters for different water models and corresponding binding free energies (kcal/mol).

Well-Tempered MetaDynamics of TRY-BEN

The optimal parameter selection for simulations depends on the specific water model employed. Models with higher viscosity, such as TIP4P, typically require a stronger bias potential to achieve effective sampling. Additionally, WT-MetaD is generally less affected by dissipative effects compared to SMD, which is more sensitive to kinetic factors. Following parameter optimization, all simulations reached convergence at $1.5 \mu\text{s}$, allowing for successful reconstruction of the PMF profiles. The resulting PMFs are presented in Figure 9.1, while a summary of the simulation parameters and outcomes is provided in Table 9.1.

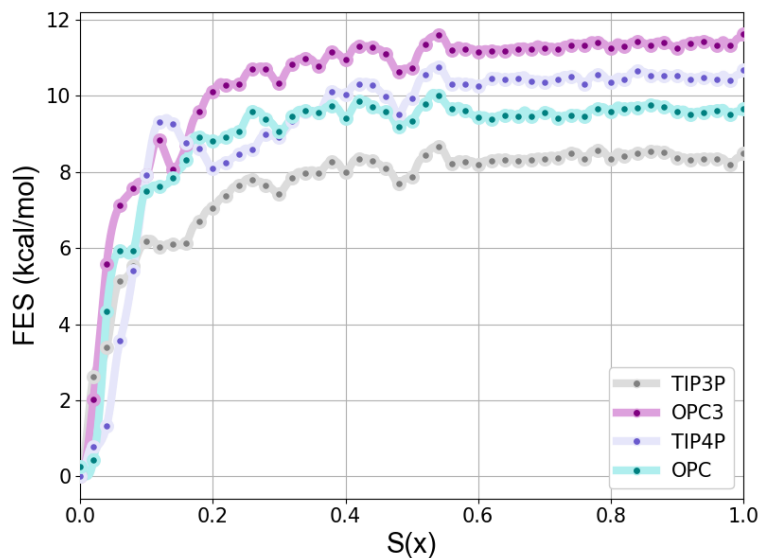


Figure 9.1: PMFs from WT-MetaD simulations with different water models.

Beginning with the simplest water model, TIP3P, the results diverge from those obtained using our nonequilibrium protocol. Specifically, the free energy estimate

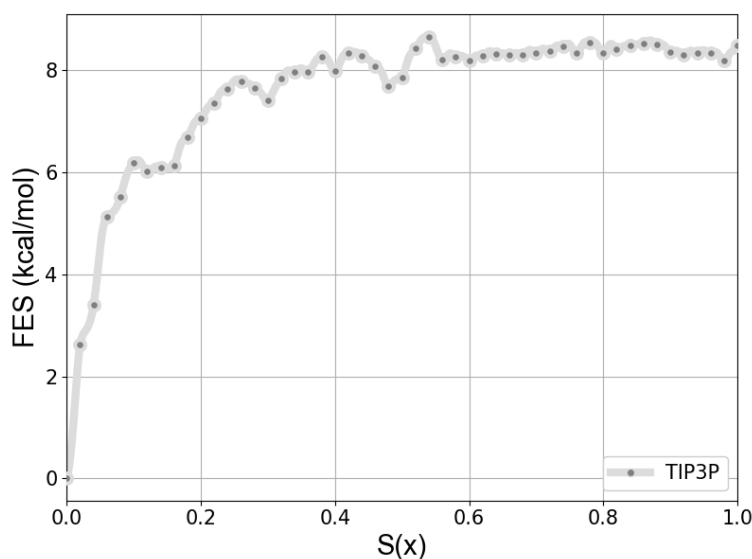


Figure 9.2: PMF from WT-MetaD in TIP3P with the same simulation parameters used for the other water models.

derived from WT-MetaD shows weaker agreement with experimental data. This observation supports our earlier hypothesis: although TIP3P may not be the most accurate model in general, it proves to be particularly well-suited for non-equilibrium SMD simulations.

Although OPC3 is a 3-point water model, it is clear that the simulation parameters optimized for TIP3P are not directly transferable. In terms of physical behavior, OPC3 more closely resembles higher-viscosity models like TIP4P and OPC. While our nonequilibrium protocol produced comparable results for both TIP3P and OPC3, the latter demonstrates superior performance in this context. This finding is consistent with recommendations from the AMBER development team and reports in the literature. Notably, OPC3 WT-MetaD simulations yielded free energy estimates more closely aligned with experimental values.

In line with our previous findings, TIP4P (despite its limitations in SMD simulations) demonstrates clear advantages over TIP3P in the context of WT-MetaD. While TIP4P may not consistently perform well with nonequilibrium estimators, it proves more effective in methods that are less sensitive to kinetic effects. This suggests that more sophisticated water models are particularly beneficial in equilibrium-

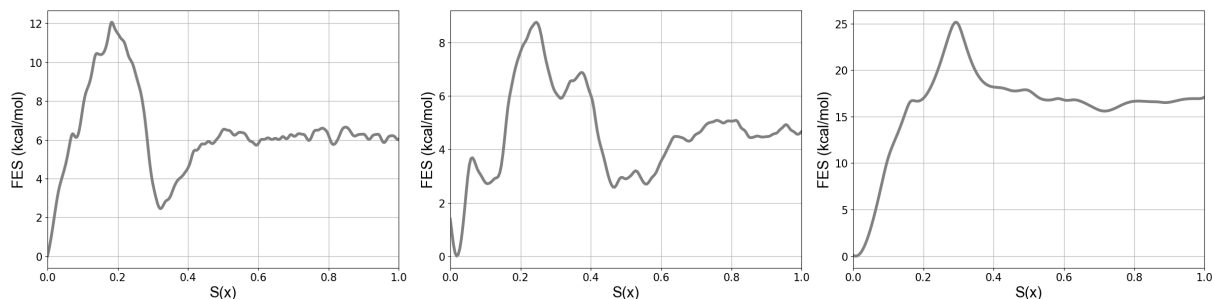


Figure 9.3: PMFs from Well-Tempered MetaD: a) With the harmonic restraint centered in $Z=0.05 \text{ nm}^2$ for the T4 Lysozyme system with Path 1 defined with ligand atoms only. b) With the harmonic restraint centered in $Z=0.05 \text{ nm}^2$ for the T4 Lysozyme system with Path 1 defined with ligand and pocket atoms. c) With the harmonic restraint centered in $Z=0.05 \text{ nm}^2$ for the T4 Lysozyme system, using Path 2 defined exclusively with ligand atoms.

like approaches. Similarly, OPC also produced free energy estimates in WT-MetaD that closely matched experimental values, further highlighting the strengths of advanced water models in this simulation framework.

Lastly, for completeness, the PMF profile generated using the TIP3P water model, employing the same simulation parameters as those used for the other water models, is also shown in Figure 9.2.

Well-Tempered MetaDynamics of T4 Lysozyme

Here the WT-MetaD results of T4 Lysozyme with PCVs and distinct reference paths are reported. For all simulations, convergence is achieved at $1 \mu\text{s}$ and PMFs are successfully reconstructed.

As shown in Figures 9.3, the discrepancies in binding free energy observed in SMD simulations are mirrored in the results from the WT-MetaD analysis. This consistency indicates that the issue does not stem from the nonequilibrium protocol itself, but is more likely related to the choice of PCVs and the construction of the reference path. Additionally, Figure 9.4 illustrates that using Path 2, defined solely by ligand atoms, leads to a system behavior similar to that seen in SMD, where it becomes trapped on the protein surface around $S(x)$ values of 39-40. This emphasizes the importance of including pocket atoms in the path definition to more accurately

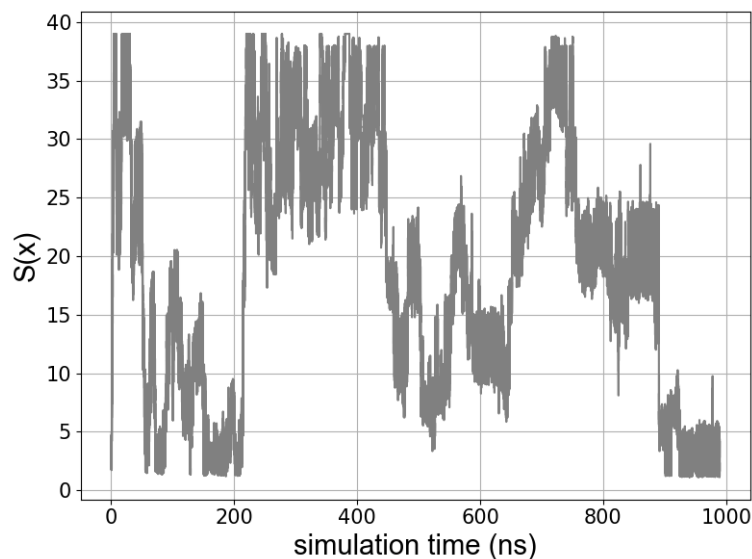


Figure 9.4: $S(X)$ as a function of the simulation time during WT-MetaD with the harmonic restraint centered in $Z=0.05 \text{ nm}^2$ for the T4 Lysozyme system, using Path 2 defined exclusively with ligand atoms.

represent system dynamics. Taken together, the WT-MetaD results validate the optimized choice of both the reference path and the PCVs. Furthermore, they underscore the inherent complexity of modeling deeply buried binding pockets, a challenge that is evident in both equilibrium and nonequilibrium simulations.

9.2 Expand the applicability of nonequilibrium binding free energy estimation: from protein to RNA

9.2.1 Additional results for RNA-ligand systems

In this section, additional results of RNA-ligand systems are reported. Particularly: the Debye-Hückel interaction energy profiles, the RMSF analysis, and the selected atoms for the course-grain like unbinding path.

9.2.2 Debye-Hückel interaction energy profile

Figure 9.5 illustrates the Debye-Hückel interaction energy profiles recorded during the ABMD simulations. As the ligands dissociate, a clear decrease in interaction energy is observed, reaching final values of 1.73 kcal/mol for the cognate ligand and 1.14 kcal/mol for the synthetic counterpart.

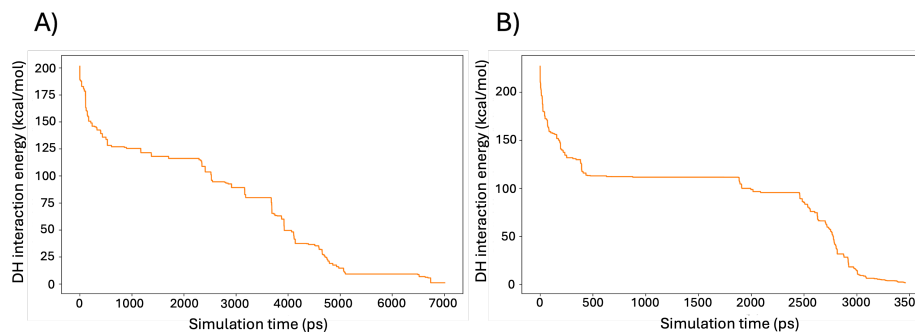


Figure 9.5: Debye-Hückel interaction energy profile of the cognate (A) and synthetic (B) ligands over simulation time

9.2.3 Analysis of PCV and Path definition

To determine suitable atoms for structural alignment, we performed three independent 100 ns unbiased MD simulations and aggregated the data to compute residue-level root-mean-square fluctuations (RMSF). Residues exhibiting RMSF values below 1.8 Å were selected for further consideration (Figure 9.6), and a subset of their atoms was used for alignment. The atoms chosen to define the coarse-grained unbinding pathway are depicted in Figure 9.7. Specifically, alignment of the RNA structure was based on phosphorus (*P*) and C1 carbon atoms of the backbone. Notably, this strategy mirrors the selection principles adopted for the protein system.

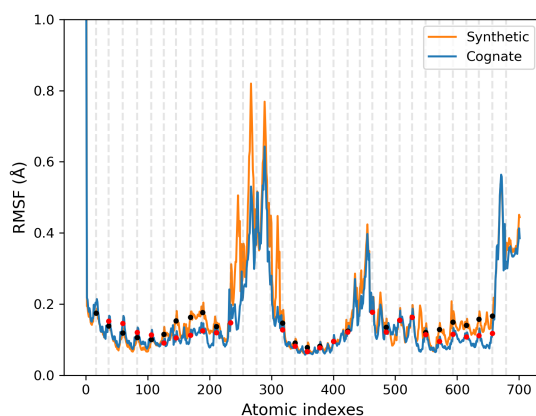


Figure 9.6: Root Mean Square Fluctuation analysis for cognate (blue) and synthetic (orange) ligands calculated across three combined 100 ns unbiased trajectories. Black and red dots represent phosphate atoms of residues with $\text{RMSF} < 0.18 \text{ \AA}$, included in the unbinding pathway.

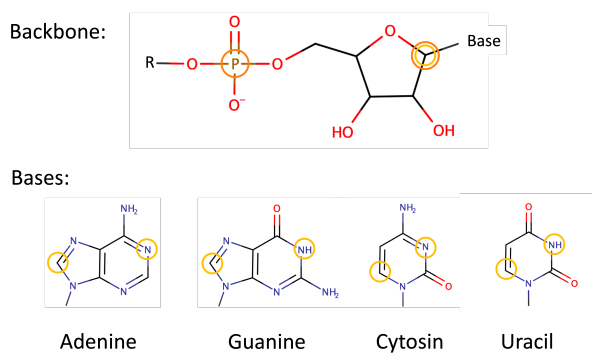


Figure 9.7: Schematic representation of the selected atoms for the course-grain like unbinding path. It is represented in orange the atom selection for the alignment and in yellow the ones for the RMSD calculation

Chapter 10

Conclusions

This thesis has investigated the application and development of binding free energy methods in the context of computational drug discovery, with a particular focus on nonequilibrium approaches and the use of path collective variables. Through a combination of theoretical analysis, methodological innovation, and practical case studies, this work contributes to the growing body of knowledge aimed at enhancing the accuracy, efficiency, and applicability of free energy simulations within the drug development pipeline.

We began by presenting a comprehensive review of binding free energy techniques, outlining the theoretical foundations of molecular dynamics simulations, enhanced sampling strategies, and equilibrium and nonequilibrium free energy estimators. This foundational review, recently published in an *Account* article, synthesized recent advances in the field and highlighted key methodological challenges that informed the direction of our research.

Building upon this foundation, we designed and implemented a novel computational pipeline that leverages nonequilibrium estimators, PCVs, and enhanced sampling techniques. This semi-automated framework was developed with a focus on scalability and usability, featuring straightforward parallelization and convergence diagnostics, making it a practical tool for broader adoption in drug discovery efforts. We validated this workflow on standard benchmark systems, such as host-guest

complexes and the TRY-BEN protein-ligand system. These studies demonstrated that our nonequilibrium approach achieves robust accuracy and performance, supporting its utility as a valuable alternative to computationally intensive equilibrium methods.

A significant insight from our investigations is the central role of dissipation in nonequilibrium simulations. In response, we proposed practical guidelines for reducing dissipation, with particular attention to the choice of water models and the optimal parametrization of PCVs. These considerations are essential for maximizing the reliability and convergence of free energy estimates in realistic biomolecular systems.

Furthermore, we extended our computational strategy to complex and pharmaceutically relevant targets, such as the Abl-Gleevec kinase-inhibitor complex and RNA-ligand systems. These case studies required specific adaptations of the pipeline, illustrating its flexibility and broader applicability beyond model systems.

Nevertheless, despite notable advances in methodologies, algorithms, and computing power, several open challenges continue to hinder the full potential of these techniques in routine applications. One of the primary obstacles remains adequate sampling of rare events, such as protein-ligand binding and unbinding, which often involve high energy barriers and long timescales. Standard MD approaches are limited by quasi-non-ergodicity, requiring the use of enhanced sampling methods, whose effectiveness is often constrained by the choice of appropriate CVs. The development of CVs that can accurately capture the relevant slow degrees of freedom is both critical and nontrivial. Promisingly, recent work has explored machine learning-based CVs, offering new strategies to improve sampling efficiency in high-dimensional systems.

A second key challenge lies in the accuracy of the FFs. Conventional FFs may lack the flexibility to describe polarization effects and subtle chemical changes, particularly for systems involving charged or flexible ligands. This limitation becomes even more pronounced when non-standard atoms or moieties are involved, as standard FFs may not provide reliable parameters for such cases, potentially compromising

the accuracy of the results. In response, machine learning-based force fields, such as neural network potentials, have emerged as a compelling direction for refining the energy landscape representation. The integration of such models into free energy workflows offers an opportunity to improve the accuracy of simulation results, though their adoption still requires further validation and standardization.

A third important limitation linked to MD simulations (whether standard or enhanced) is the assumption of fixed protonation states for ionizable groups. In practice, the protonation state of most biomolecules is not static, but fluctuates over time in response to the pH of the surrounding environment. For many systems, pH-induced protonation changes are essential to their biological function, directly affecting conformational dynamics and molecular interactions. However, conventional MD simulations neglect this aspect by assigning fixed protonation states, typically based on the most probable state at the starting pH, thereby limiting the realism of the simulation. This simplification is particularly critical when studying processes such as protein-ligand binding, where dynamic protonation changes can significantly alter binding mechanisms and free energy profiles. To address this challenge and improve the chemical realism of biomolecular simulations, efforts have been directed toward the development of constant-pH MD protocols. One promising approach involves the use of λ -dynamics combined with fast multipole method electrostatics, enabling the on-the-fly sampling of protonation states during the simulation. While computationally demanding, these methods represent a crucial step toward more accurate and predictive modeling of biochemical systems under physiological conditions.

Another, often underestimated, challenge is convergence. Regardless of the method employed (be it equilibrium or nonequilibrium free energy estimators) convergence is essential for generating reproducible and statistically robust results. In practice, this remains a critical bottleneck, particularly for nonequilibrium methods such as those based on JE. Systematic approaches for assessing convergence, including bidirectional estimators and statistical error analyses, must therefore become standard practice in computational workflows.

Finally, the balance between physical and alchemical methods presents both an

opportunity and a challenge. While alchemical methods are powerful for large-scale affinity predictions (e.g., in hit-to-lead campaigns), they are limited in scope due to convergence issues with large or highly polar ligands. Conversely, physical, path-based approaches provide mechanistic insight but rely heavily on prior knowledge of suitable CVs. Emerging hybrid methods, such as the Alchemical Transfer Method, seek to bridge these paradigms, but further research is needed to fully integrate such frameworks in the drug discovery context.

Overall, while the free energy landscape has seen substantial methodological and theoretical progress, the path to routine, high-confidence application in drug discovery remains complex. Overcoming current limitations will depend on a synergistic effort involving improved sampling techniques, next-generation FFs, robust statistical estimators, and the incorporation of machine learning at various levels of the simulation workflow. Addressing these challenges will ultimately enable the broader and more reliable use of MD-based free energy methods in real-world drug development, accelerating the transition from computational predictions to therapeutic impact.

In conclusion, the research presented in this thesis makes a substantial contribution to the advancement of computational methodologies for drug discovery, with a particular focus on the application and development of nonequilibrium free-energy approaches. The results obtained in the published works demonstrate that these methods can provide accurate, robust, and computationally efficient estimates of protein–ligand binding free energies for biologically and pharmacologically relevant systems. By validating the proposed strategies against experimental data, this work highlights the practical predictive power of nonequilibrium techniques, reinforcing their relevance for real-world drug discovery applications.

Beyond the core methodological developments, the inclusion of complementary projects, such as the development of NanoShaperWeb and the investigation of ligand residence time estimation, strengthens the interdisciplinary scope of this thesis. These side projects are closely connected to the central theme, as they address key aspects of molecular recognition, binding-site characterization, and kinetic behavior that are essential for a comprehensive understanding of protein–ligand interactions.

Together, the methodologies and tools presented here support the ongoing transformation of *in-silico* drug design and are expected to serve as valuable assets for future studies in both academic and pharmaceutical research settings.

Publications

- [P1] Ghidini, Alessia*, Serra, Eleonora*, Decherchi S, Cavalli A. Bidirectional path-based non-equilibrium simulations for binding free energy. *Molecular Physics*. 2024:e2374465.
- [P2] Serra, Eleonora*, Ghidini, Alessia*, Decherchi S, Cavalli A. Nonequilibrium Binding Free Energy Simulations: Minimizing Dissipation. *Journal of Chemical Theory and Computation*. 2025.
- [P3] Serra, Eleonora, Ghidini, Alessia, Aguti R, Bernetti M, Decherchi S, Cavalli A. Path-Based Nonequilibrium Binding Free Energy Estimation, from Protein–Ligand to RNA-Ligand Binding. *Journal of Chemical Information and Modeling*. 2025.
- [P4] Abate, Carlo*, Serra, Eleonora*, Rocchia W, Decherchi S. NanoShaperWeb: Molecular Surface and Pocket Detection Made Visual. *Journal of Chemical Information and Modeling*. 2025.
- [P5] Ghidini, Alessia*, Serra, Eleonora*, Cavalli A. On Free Energy Calculations in Drug Discovery. *Accounts of Chemical Research*. 2025:5287-300.

*These authors contributed equally to the work.

Bibliography

- [1] Singh N, Vayer P, Tanwar S, Poyet JL, Tsaioun K, Villoutreix BO. Drug discovery and development: introduction to the general public and patient groups. *Frontiers in Drug Discovery*. 2023;3:1201419.
- [2] Berdigaliyev N, Aljofan M. An overview of drug discovery and development. *Future medicinal chemistry*. 2020;12(10):939-47.
- [3] Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature reviews Drug discovery*. 2010;9(3):203-14.
- [4] Decherchi S, Cavalli A. Thermodynamics and kinetics of drug-target binding by molecular simulation. *Chemical Reviews*. 2020;120(23):12788-833.
- [5] Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. *Nature*. 2023;616(7958):673-85.
- [6] Gellman SH. Introduction: molecular recognition. *Chemical reviews*. 1997;97(5):1231-2.
- [7] Du X, Li Y, Xia YL, Ai SM, Liang J, Sang P, et al. Insights into protein–ligand interactions: mechanisms, models, and methods. *International journal of molecular sciences*. 2016;17(2):144.
- [8] Wong E. Cells: molecules and mechanisms. Axolotl Academic Publishing; 2009.
- [9] Guyton AC. Text book of medical physiology. China; 2006.

- [10] Verdine GL, Walensky LD. The challenge of drugging undruggable targets in cancer: lessons learned from targeting BCL-2 family members. *Clinical cancer research*. 2007;13(24):7264-70.
- [11] De Vivo M, Masetti M, Bottegoni G, Cavalli A. Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry*. 2016;59(9):4035-61.
- [12] Nunes-Alves A, Zuckerman DM, Arantes GM. Escape of a small molecule from inside T4 lysozyme by multiple pathways. *Biophysical journal*. 2018;114(5):1058-66.
- [13] Copeland RA, Pompliano DL, Meek TD. Drug–target residence time and its implications for lead optimization. *Nature reviews Drug discovery*. 2006;5(9):730-9.
- [14] Fischer E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft*. 1894;27(3):2985-93.
- [15] Koshland Jr DE. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences*. 1958;44(2):98-104.
- [16] Csermely P, Palotai R, Nussinov R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in biochemical sciences*. 2010;35(10):539-46.
- [17] Salahudeen MS, Nishtala PS. An overview of pharmacodynamic modelling, ligand-binding approach and its application in clinical practice. *Saudi Pharmaceutical Journal*. 2017;25(2):165-75.
- [18] McCammon JA. Theory of biomolecular recognition. *Current opinion in structural biology*. 1998;8(2):245-9.
- [19] Copeland RA. The drug–target residence time model: a 10-year retrospective. *Nature Reviews Drug Discovery*. 2016;15(2):87-95.

- [20] Tang Z, Roberts CC, Chia-en AC. Understanding ligand-receptor non-covalent binding kinetics using molecular modeling. *Frontiers in bioscience* (Landmark edition). 2017;22:960.
- [21] Gedde UW, Helmholtz F. *Essential classical thermodynamics*. Springer; 2020.
- [22] Kukol A, et al. *Molecular modeling of proteins*. vol. 443. Springer; 2008.
- [23] Gedde UW, Gedde UW. *Gibbs and Helmholtz Free Energies. Essential Classical Thermodynamics*. 2020:21-4.
- [24] Bronowska AK. Thermodynamics of ligand-protein interactions: implications for molecular design. In: *Thermodynamics-Interaction Studies-Solids, Liquids and Gases*. IntechOpen; 2011. .
- [25] Chang CeA, Chen W, Gilson MK. Ligand configurational entropy and protein binding. *Proceedings of the National Academy of Sciences*. 2007;104(5):1534-9.
- [26] MacRaild CA, Daranas AH, Bronowska A, Homans SW. Global changes in local protein dynamics reduce the entropic cost of carbohydrate binding in the arabinose-binding protein. *Journal of molecular biology*. 2007;368(3):822-32.
- [27] Bernetti M, Masetti M, Rocchia W, Cavalli A. Kinetics of drug binding and residence time. *Annual review of physical chemistry*. 2019;70:143-71.
- [28] Bongrand P. Ligand-receptor interactions. *Reports on progress in Physics*. 1999;62(6):921.
- [29] Van't Hoff JH, Hoff JH. *Etudes de dynamique chimique*. F. Muller & Company; 1884.
- [30] Arrhenius S. Über die Dissociationswärme und den Einfluss der Temperatur auf den Dissociationsgrad der Elektrolyte. *Zeitschrift für physikalische Chemie*. 1889;4(1):96-116.
- [31] Goldberger ML, Watson KM. *Collision theory*. Courier Corporation; 2004.

- [32] Truhlar DG, Garrett BC, Klippenstein SJ. Current status of transition-state theory. *The Journal of physical chemistry*. 1996;100(31):12771-800.
- [33] Laidler KJ, King MC. The development of transition-state theory. *J phys Chem*. 1983;87(15):2657-64.
- [34] Eyring H. The activated complex in chemical reactions. *The Journal of Chemical Physics*. 1935;3(2):107-15.
- [35] Wang J, Do HN, Koirala K, Miao Y. Predicting biomolecular binding kinetics: A review. *Journal of Chemical Theory and Computation*. 2023;19(8):2135-48.
- [36] Nunes-Alves A, Kokh DB, Wade RC. Recent progress in molecular simulation methods for drug binding kinetics. *Current Opinion in Structural Biology*. 2020;64:126-33.
- [37] Swinney DC. The role of binding kinetics in therapeutically useful drug action. *Current opinion in drug discovery & development*. 2009;12(1):31-9.
- [38] Callegari D, Lodola A, Pala D, Rivara S, Mor M, Rizzi A, et al. Metadynamics simulations distinguish short-and long-residence-time inhibitors of cyclin-dependent kinase 8. *Journal of chemical information and modeling*. 2017;57(2):159-69.
- [39] Tummino PJ, Copeland RA. Residence time of receptor- ligand complexes and its effect on biological function. *Biochemistry*. 2008;47(20):5481-92.
- [40] Sohraby F, Nunes-Alves A. Advances in computational methods for ligand binding kinetics. *Trends in Biochemical Sciences*. 2023;48(5):437-49.
- [41] Ahmad K, Rizzi A, Capelli R, Mandelli D, Lyu W, Carloni P. Enhanced-sampling simulations for the estimation of ligand binding kinetics: current status and perspective. *Frontiers in molecular biosciences*. 2022;9:899805.
- [42] Tiwary P, Limongelli V, Salvalaglio M, Parrinello M. Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proceedings of the National Academy of Sciences*. 2015;112(5):E386-91.

- [43] Wolf S. Predicting protein–ligand binding and unbinding kinetics with biased MD simulations and coarse-graining of dynamics: Current state and challenges. *Journal of Chemical Information and Modeling*. 2023;63(10):2902-10.
- [44] Schuetz DA, de Witte WEA, Wong YC, Knasmueller B, Richter L, Kokh DB, et al. Kinetics for Drug Discovery: an industry-driven effort to target drug residence time. *Drug discovery today*. 2017;22(6):896-911.
- [45] Leach AR. *Molecular modelling: principles and applications*. Pearson education; 2001.
- [46] Tuckerman ME. *Statistical mechanics: theory and molecular simulation*. Oxford university press; 2023.
- [47] Goodstein DL. *States of matter*. Courier Corporation; 1985.
- [48] Singh N. On the foundations of statistical mechanics: a brief review. *Modern Physics Letters B*. 2013;27(06):1330003.
- [49] Allen MP, Tildesley DJ. *Computer simulation of liquids*. Oxford university press; 2017.
- [50] Poltev V. *Molecular mechanics: principles, history, and current status*. *Handbook of Computational Chemistry*; Leszczynski, J, Ed; Springer: Dordrecht, The Netherlands. 2015.
- [51] Born M, Heisenberg W. Zur quantentheorie der molekeln. *Original Scientific Papers Wissenschaftliche Originalarbeiten*. 1985:216-46.
- [52] Vanommeslaeghe K, Guvench O, et al. Molecular mechanics. *Current pharmaceutical design*. 2014;20(20):3281-92.
- [53] Hollingsworth SA, Dror RO. Molecular dynamics simulation for all. *Neuron*. 2018;99(6):1129-43.
- [54] Tuckerman ME, Martyna GJ. *Understanding modern molecular dynamics: Techniques and applications*. ACS Publications; 2000.

- [55] Olivei A. Surface imaging by scanning-laser-induced voltage. *Journal of Physics D: Applied Physics*. 1976;9(2):183.
- [56] Young P. The leapfrog method and other “symplectic” algorithms for integrating Newton’s laws of motion. *Lecture notes in University of california, santa cruz*. 2014.
- [57] Frenkel D, Smit B. *Understanding molecular simulation: from algorithms to applications*. Academic press. 2001;1.
- [58] Ryckaert JP, Ciccotti G, Berendsen HJ. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of computational physics*. 1977;23(3):327-41.
- [59] Hess B, Bekker H, Berendsen HJ, Fraaije JG. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry*. 1997;18(12):1463-72.
- [60] Miyamoto S, Kollman PA. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of computational chemistry*. 1992;13(8):952-62.
- [61] Thomas ME, Vidya R, Thomas J, Ahmad Z. Theoretical Aspects of XLPE-Based Blends and Nanocomposites. *Crosslinkable Polyethylene Based Blends and Nanocomposites*. 2021:299-319.
- [62] Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*. 1995;117(19):5179-97.
- [63] Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation*. 2015;11(8):3696-713.
- [64] Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of

- organic liquids. *Journal of the american chemical society*. 1996;118(45):11225-36.
- [65] Huang J, MacKerell Jr AD. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of computational chemistry*. 2013;34(25):2135-45.
 - [66] Patel S, Brooks III CL. CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *Journal of computational chemistry*. 2004;25(1):1-16.
 - [67] Schuler LD, Daura X, Van Gunsteren WF. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *Journal of computational chemistry*. 2001;22(11):1205-18.
 - [68] Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, et al. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*. 1984;106(3):765-84.
 - [69] Hagler A, Huler E, Lifson S. Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *Journal of the American Chemical Society*. 1974;96(17):5319-27.
 - [70] Weiner SJ, Kollman PA, Nguyen DT, Case DA. An all atom force field for simulations of proteins and nucleic acids. *Journal of computational chemistry*. 1986;7(2):230-52.
 - [71] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*. 2006;65(3):712-25.
 - [72] Wang ZX, Zhang W, Wu C, Lei H, Cieplak P, Duan Y. Strike a balance: optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides. *Journal of computational chemistry*. 2006;27(6):781-90.

- [73] Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. *Journal of computational chemistry*. 2004;25(9):1157-74.
- [74] He X, Man VH, Yang W, Lee TS, Wang J. A fast and high-quality charge model for the next generation general AMBER force field. *The Journal of Chemical Physics*. 2020;153(11).
- [75] Bayly CI, Cieplak P, Cornell W, Kollman PA. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry*. 1993;97(40):10269-80.
- [76] Jakalian A, Bush BL, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *Journal of computational chemistry*. 2000;21(2):132-46.
- [77] Connolly ML. Analytical molecular surface calculation. *Journal of applied crystallography*. 1983;16(5):548-58.
- [78] Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S. Protein Data Bank (PDB): the single global macromolecular structure archive. *Protein crystallography: methods and protocols*. 2017:627-41.
- [79] Lemkul JA. Introductory tutorials for simulating protein dynamics with GROMACS. *The Journal of Physical Chemistry B*. 2024;128(39):9418-35.
- [80] Kamberaj H. *Molecular dynamics simulations in statistical physics: theory and applications*. Springer; 2020.
- [81] Varma S, Chiu SW, Jakobsson E. The influence of amino acid protonation states on molecular dynamics simulations of the bacterial porin OmpF. *Biophysical journal*. 2006;90(1):112-23.
- [82] Tsai CC, Yue Z, Shen J. How electrostatic coupling enables conformational plasticity in a tyrosine kinase. *Journal of the American Chemical Society*. 2019;141(38):15092-101.

- [83] Harris RC, Shen J. GPU-accelerated implementation of continuous constant pH molecular dynamics in amber: p K a predictions with single-pH simulations. *Journal of chemical information and modeling*. 2019;59(11):4821-32.
- [84] Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS: fast, flexible, and free. *Journal of computational chemistry*. 2005;26(16):1701-18.
- [85] Lingenheil M, Denschlag R, Reichold R, Tavan P. The “hot-solvent/cold-solute” problem revisited. *Journal of Chemical Theory and Computation*. 2008;4(8):1293-306.
- [86] Berendsen HJ, Postma Jv, Van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*. 1984;81(8):3684-90.
- [87] Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. *The Journal of chemical physics*. 2007;126(1).
- [88] Wassenaar TA, Mark AE. The effect of box shape on the dynamic properties of proteins simulated under periodic boundary conditions. *Journal of computational chemistry*. 2006;27(3):316-25.
- [89] Ewald PP. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der physik*. 1921;369(3):253-87.
- [90] Darden T, York D, Pedersen L. Particle mesh Ewald: An $N \log(N)$ method for Ewald sums in large systems. *The Journal of chemical physics*. 1993;98(12):10089-92.
- [91] Cozzini P, Fornabaio M, Marabotti A, Abraham DJ, Kellogg GE, Mozzarelli A. Free energy of ligand binding to protein: evaluation of the contribution of water molecules by computational methods. *Current medicinal chemistry*. 2004;11(23):3093-118.
- [92] Wang L, Berne B, Friesner R. Ligand binding to protein-binding pockets with wet and dry regions. *Proceedings of the National Academy of Sciences*. 2011;108(4):1326-30.

- [93] Nguyen TT, Viet MH, Li MS, et al. Effects of water models on binding affinity: evidence from all-atom simulation of binding of tamiflu to A/H5N1 neuraminidase. *The Scientific World Journal*. 2014;2014.
- [94] Ando T. Shear viscosity of OPC and OPC3 water models. *The Journal of Chemical Physics*. 2023;159(10).
- [95] Decherchi S, Masetti M, Vyalov I, Rocchia W. Implicit solvent methods for free energy estimation. *European journal of medicinal chemistry*. 2015;91:27-42.
- [96] Kadaoluwa Pathirannahalage SP, Meftahi N, Elbourne A, Weiss AC, McConville CF, Padua A, et al. Systematic comparison of the structural and dynamic properties of commonly used water models for molecular dynamics simulations. *Journal of Chemical Information and Modeling*. 2021;61(9):4521-36.
- [97] Onufriev AV, Izadi S. Water models for biomolecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. 2018;8(2):e1347.
- [98] Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*. 1983;79(2):926-35.
- [99] Schmitt UW, Voth GA. The computer simulation of proton transport in water. *The Journal of chemical physics*. 1999;111(20):9361-81.
- [100] Izadi S, Onufriev AV. Accuracy limit of rigid 3-point water models. *The Journal of chemical physics*. 2016;145(7).
- [101] Jorgensen WL, Madura JD. Temperature and size dependence for Monte Carlo simulations of TIP4P water. *Molecular Physics*. 1985;56(6):1381-92.
- [102] González MA, Abascal JL. A flexible model for water based on TIP4P/2005. *The Journal of chemical physics*. 2011;135(22).
- [103] Aguilar-Pineda JA, González-Melchor M. Influence of the Water Model on the Structure and Interactions of the GPR40 Protein with the Lipid Mem-

- brane and the Solvent: Rigid versus Flexible Water Models. *Journal of Chemical Theory and Computation*. 2024;20(14):6369-87.
- [104] Izadi S, Anandakrishnan R, Onufriev AV. Building water models: a different approach. *The journal of physical chemistry letters*. 2014;5(21):3863-71.
 - [105] Chipot C, Pohorille A. *Free energy calculations*. vol. 86. Berlin Heidelberg: Springer; 2007.
 - [106] Bernardi RC, Melo MC, Schulten K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA)-General Subjects*. 2015;1850(5):872-7.
 - [107] Shaw DE, Deneroff MM, Dror RO, Kuskin JS, Larson RH, Salmon JK, et al. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*. 2008;51(7):91-7.
 - [108] Hénin J, Lelièvre T, Shirts MR, Valsson O, Delemotte L. Enhanced sampling methods for molecular dynamics simulations. *arXiv preprint arXiv:220204164*. 2022.
 - [109] De Donder T. Transformations physiques et chimiques des systèmes de Gibbs. *Bulletin de la Classe des sciences Académie royale de Belgique*. 1920;1920:315-28.
 - [110] Rocci A. Celebrating the birth of De Donder’s chemical affinity (1922–2022): from the uncompensated heat to his Ave Maria. *Foundations of Chemistry*. 2024;26(1):37-73.
 - [111] Kirkwood JG. Statistical mechanics of fluid mixtures. *The Journal of chemical physics*. 1935;3(5):300-13.
 - [112] Chipot C. Free energy methods for the description of molecular processes. *Annual Review of Biophysics*. 2023;52(1):113-38.
 - [113] Wong KY, York DM. Exact relation between potential of mean force and free-energy profile. *Journal of chemical theory and computation*. 2012;8(11):3998-4003.

- [114] Cavalli A, Spitaleri A, Saladino G, Gervasio FL. Investigating drug–target association and dissociation mechanisms using metadynamics-based algorithms. *Accounts of chemical research*. 2015;48(2):277-85.
- [115] Branduardi D, Gervasio FL, Parrinello M. From A to B in free energy space. *The Journal of chemical physics*. 2007;126(5).
- [116] Hovan L, Comitani F, Gervasio FL. Defining an optimal metric for the path collective variables. *Journal of Chemical Theory and Computation*. 2018;15(1):25-32.
- [117] Bonomi M, Branduardi D. PLUMED Tutorial: A portable plugin for free-energy calculations with molecular dynamics. <http://sites.google.com/site/plumedtutorial2010/>. 2010.
- [118] Torrie GM, Valleau JP. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of computational physics*. 1977;23(2):187-99.
- [119] Zhu F, Hummer G. Convergence and error estimation in free energy calculations using the weighted histogram analysis method. *Journal of computational chemistry*. 2012;33(4):453-65.
- [120] Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of computational chemistry*. 1992;13(8):1011-21.
- [121] Laio A, Parrinello M. Escaping free-energy minima. *Proceedings of the national academy of sciences*. 2002;99(20):12562-6.
- [122] Ensing B, De Vivo M, Liu Z, Moore P, Klein ML. Metadynamics as a tool for exploring free energy landscapes of chemical reactions. *Accounts of chemical research*. 2006;39(2):73-81.
- [123] Laio A, Rodriguez-Fortea A, Gervasio FL, Ceccarelli M, Parrinello M. Assessing the accuracy of metadynamics. *The journal of physical chemistry B*. 2005;109(14):6714-21.

- [124] Barducci A, Bussi G, Parrinello M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters*. 2008;100(2):020603.
- [125] Dama JF, Parrinello M, Voth GA. Well-tempered metadynamics converges asymptotically. *Physical review letters*. 2014;112(24):240602.
- [126] Raiteri P, Laio A, Gervasio FL, Micheletti C, Parrinello M. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *The journal of physical chemistry B*. 2006;110(8):3533-9.
- [127] Branduardi D, Bussi G, Parrinello M. Metadynamics with adaptive Gaussians. *Journal of chemical theory and computation*. 2012;8(7):2247-54.
- [128] Gervasio FL, Laio A, Parrinello M. Flexible docking in solution using metadynamics. *Journal of the American Chemical Society*. 2005;127(8):2600-7.
- [129] Izrailev S, Stepaniants S, Isralewitz B, Kosztin D, Lu H, Molnar F, et al. Steered molecular dynamics. In: *Computational Molecular Dynamics: Challenges, Methods, Ideas: Proceedings of the 2nd International Symposium on Algorithms for Macromolecular Modelling*, Berlin, May 21–24, 1997. Springer; 1999. p. 39-65.
- [130] Isralewitz B, Gao M, Schulten K. Steered molecular dynamics and mechanical functions of proteins. *Current opinion in structural biology*. 2001;11(2):224-30.
- [131] Lu H, Schulten K. Steered molecular dynamics simulations of force-induced protein domain unfolding. *Proteins: Structure, Function, and Bioinformatics*. 1999;35(4):453-63.
- [132] Izrailev S, Stepaniants S, Balsera M, Oono Y, Schulten K. Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophysical journal*. 1997;72(4):1568-81.
- [133] Park S, Schulten K. Calculating potentials of mean force from steered molecular dynamics simulations. *The Journal of chemical physics*. 2004;120(13):5946-61.

- [134] Colizzi F, Perozzo R, Scapozza L, Recanatini M, Cavalli A. Single-molecule pulling simulations can discern active from inactive enzyme inhibitors. *Journal of the American Chemical Society*. 2010;132(21):7361-71.
- [135] Marchi M, Ballone P. Adiabatic bias molecular dynamics: a method to navigate the conformational space of complex molecular systems. *The Journal of chemical physics*. 1999;110(8):3697-702.
- [136] Christ CD, Mark AE, Van Gunsteren WF. Basic ingredients of free energy calculations: a review. *Journal of computational chemistry*. 2010;31(8):1569-82.
- [137] Trzesniak D, Kunz APE, van Gunsteren WF. A comparison of methods to compute the potential of mean force. *ChemPhysChem*. 2007;8(1):162-9.
- [138] Van Gunsteren WF, Daura X, Mark AE. Computation of free energy. *Helvetica Chimica Acta*. 2002;85(10):3113-29.
- [139] Pohorille A, Jarzynski C, Chipot C. Good practices in free-energy calculations. *The Journal of Physical Chemistry B*. 2010;114(32):10235-53.
- [140] Zwanzig RW. High-temperature equation of state by a perturbation method. I. Nonpolar gases. *The Journal of Chemical Physics*. 1954;22(8):1420-6.
- [141] Bennett CH. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*. 1976;22(2):245-68.
- [142] Shirts MR, Bair E, Hooker G, Pande VS. Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods. *Physical review letters*. 2003;91(14):140601.
- [143] Ciccotti G, Rondoni L. Jarzynski on work and free energy relations: The case of variable volume. *American Institute of Chemical Engineers Journals*. 2021;67(1):e17082.
- [144] Jarzynski C. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*. 1997;56(5):5018.

- [145] Jarzynski C. Nonequilibrium equality for free energy differences. *Physical Review Letters*. 1997;78(14):2690.
- [146] Procacci P. Precision and computational efficiency of nonequilibrium alchemical methods for computing free energies of solvation. II. Unidirectional estimates. *The Journal of Chemical Physics*. 2019;151(14).
- [147] Baştuğ T, Chen PC, Patra SM, Kuyucak S. Potential of mean force calculations of ligand binding to ion channels from Jarzynski’s equality and umbrella sampling. *The Journal of Chemical Physics*. 2008 04;128(15):155104.
- [148] Crooks GE. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *Journal of Statistical Physics*. 1998;90:1481-7.
- [149] Hummer G. Nonequilibrium methods for equilibrium free energy calculations. In: Chipot C, Pohorille A, editors. *Free Energy Calculations: Theory and Applications in Chemistry and Biology*. Berlin Heidelberg: Springer; 2007. p. 171-98.
- [150] Crooks GE. Path-ensemble averages in systems driven far from equilibrium. *Phys Rev E*. 2000 Mar;61:2361-6. Available from: <https://link.aps.org/doi/10.1103/PhysRevE.61.2361>.
- [151] Decherchi S, Ciccotti G, Cavalli A. Regularized Bennett and Zwanzig free energy estimators. *The Journal of Chemical Physics*. 2023 03;158(12):124101. Available from: <https://doi.org/10.1063/5.0137837>.
- [152] Procacci P. Accuracy, precision, and efficiency of nonequilibrium alchemical methods for computing free energies of solvation. I. Bidirectional approaches. *The Journal of Chemical Physics*. 2019;151(14).
- [153] Gapsys V, Yildirim A, Aldeghi M, Khalak Y, Van der Spoel D, de Groot BL. Accurate absolute free energies for ligand–protein binding based on nonequilibrium approaches. *Communications Chemistry*. 2021;4(1):61.

- [154] Procacci P, Marsili S, Barducci A, Signorini GF, Chelli R. Crooks equation for steered molecular dynamics using a Nosé-Hoover thermostat. *The Journal of chemical physics*. 2006;125(16).
- [155] Hefti FF. Requirements for a lead compound to become a clinical candidate. *BMC neuroscience*. 2008;9(Suppl 3):S7.
- [156] Jorgensen WL, Ravimohan C. Monte Carlo simulation of differences in free energies of hydration. *The Journal of chemical physics*. 1985;83(6):3050-4.
- [157] Tembre BL, Mc Cammon JA. Ligand-receptor interactions. *Computers & Chemistry*. 1984;8(4):281-3.
- [158] Cournia Z, Allen B, Sherman W. Relative binding free energy calculations in drug discovery: recent advances and practical considerations. *Journal of chemical information and modeling*. 2017;57(12):2911-37.
- [159] Jorgensen WL, Ruiz-Caro J, Tirado-Rives J, Basavapathruni A, Anderson KS, Hamilton AD. Computer-aided design of non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Bioorganic & medicinal chemistry letters*. 2006;16(3):663-7.
- [160] Jorgensen WL, Buckner JK, Boudon S, Tirado-Rives J. Efficient computation of absolute free energies of binding by computer simulations. Application to the methane dimer in water. *The Journal of chemical physics*. 1988;89(6):3742-6.
- [161] Gilson MK, Given JA, Bush BL, McCammon JA. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical journal*. 1997;72(3):1047-69.
- [162] Gumbart JC, Roux B, Chipot C. Standard binding free energies from computer simulations: What is the best strategy? *Journal of chemical theory and computation*. 2013;9(1):794-802.
- [163] Fu H, Chen H, Blazhynska M, Goulard Coderc de Lacam E, Szczepaniak F, Pavlova A, et al. Accurate determination of protein: ligand standard

- binding free energies from molecular dynamics simulations. *Nature protocols*. 2022;17(4):1114-41.
- [164] Lin YL, Roux B. Computational analysis of the binding specificity of Gleevec to Abl, c-Kit, Lck, and c-Src tyrosine kinases. *Journal of the American Chemical Society*. 2013;135(39):14741-53.
 - [165] Dokoumetzidis A, Macheras P. A century of dissolution research: from Noyes and Whitney to the biopharmaceutics classification system. *International journal of pharmaceutics*. 2006;321(1-2):1-11.
 - [166] Gobbo D, Ballone P, Decherchi S, Cavalli A. Solubility advantage of amorphous ketoprofen. Thermodynamic and kinetic aspects by molecular dynamics and free energy approaches. *Journal of Chemical Theory and Computation*. 2020;16(7):4126-40.
 - [167] Kästner J, Thiel W. Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: "Umbrella integration". *The Journal of chemical physics*. 2005;123(14).
 - [168] Rosta E, Hummer G. Free energies from dynamic weighted histogram analysis using unbiased Markov state model. *Journal of chemical theory and computation*. 2015;11(1):276-85.
 - [169] Patey G, Valleau J. A Monte Carlo method for obtaining the interionic potential of mean force in ionic solution. *The Journal of Chemical Physics*. 1975;63(6):2334-9.
 - [170] Bartels C, Karplus M. Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. *Journal of Computational Chemistry*. 1997;18(12):1450-62.
 - [171] Bigot B, Jorgensen WL. Sampling methods for Monte Carlo simulations of n-butane in dilute solution. *The Journal of Chemical Physics*. 1981;75(4):1944-52.
 - [172] Branduardi D, De Vivo M, Rega N, Barone V, Cavalli A. Methyl phosphate dianion hydrolysis in solution characterized by path collective variables cou-

- pled with DFT-based enhanced sampling simulations. *Journal of Chemical Theory and Computation*. 2011;7(3):539-43.
- [173] Lodola A, Branduardi D, De Vivo M, Capoferri L, Mor M, Piomelli D, et al. A catalytic mechanism for cysteine N-terminal nucleophile hydrolases, as revealed by free energy simulations. *PLoS One*. 2012;7(2):e32397.
- [174] Di Martino GP, Masetti M, Cavalli A, Recanatini M. Mechanistic insights into P in1 peptidyl-prolyl cis-trans isomerization from umbrella sampling simulations. *Proteins: Structure, Function, and Bioinformatics*. 2014;82(11):2943-56.
- [175] Decherchi S, Berteotti A, Bottegoni G, Rocchia W, Cavalli A. The ligand binding mechanism to purine nucleoside phosphorylase elucidated via molecular dynamics and machine learning. *Nature communications*. 2015;6(1):6155.
- [176] Maragliano L, Vanden-Eijnden E. Single-sweep methods for free energy calculations. *The Journal of chemical physics*. 2008;128(18).
- [177] Limongelli V, Bonomi M, Marinelli L, Gervasio FL, Cavalli A, Novellino E, et al. Molecular basis of cyclooxygenase enzymes (COXs) selective inhibition. *Proceedings of the National Academy of Sciences*. 2010;107(12):5411-6.
- [178] Masetti M, Cavalli A, Recanatini M, Gervasio FL. Exploring Complex Protein- Ligand Recognition Mechanisms with Coarse Metadynamics. *The journal of physical chemistry B*. 2009;113(14):4807-16.
- [179] Bernetti M, Masetti M, Recanatini M, Amaro RE, Cavalli A. An integrated Markov state model and path metadynamics approach to characterize drug binding processes. *Journal of chemical theory and computation*. 2019;15(10):5689-702.
- [180] Bertazzo M, Gobbo D, Decherchi S, Cavalli A. Machine learning and enhanced sampling simulations for computing the potential of mean force and standard binding free energy. *Journal of chemical theory and computation*. 2021;17(8):5287-300.

- [181] Lüdemann SK, Lounnas V, Wade RC. How do substrates enter and products exit the buried active site of cytochrome P450cam? 1. Random expulsion molecular dynamics investigation of ligand access channels and mechanisms. *Journal of molecular biology*. 2000;303(5):797-811.
- [182] Patel JS, Branduardi D, Masetti M, Rocchia W, Cavalli A. Insights into ligand–protein binding from local mechanical response. *Journal of Chemical Theory and Computation*. 2011;7(10):3368-78.
- [183] Patel JS, Berteotti A, Ronsisvalle S, Rocchia W, Cavalli A. Steered molecular dynamics simulations for studying protein–ligand interaction in cyclin-dependent kinase 5. *Journal of chemical information and modeling*. 2014;54(2):470-80.
- [184] Mai BK, Viet MH, Li MS. Top leads for swine influenza A/H1N1 virus revealed by steered molecular dynamics approach. *Journal of chemical information and modeling*. 2010;50(12):2236-47.
- [185] Mai BK, Li MS. Neuraminidase inhibitor R-125489—a promising drug for treating influenza virus: steered molecular dynamics approach. *Biochemical and biophysical research communications*. 2011;410(3):688-91.
- [186] Crooks GE. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*. 1999;60(3):2721.
- [187] Azimi S, Khuttan S, Wu JZ, Pal RK, Gallicchio E. Relative binding free energy calculations for ligands with diverse scaffolds with the alchemical transfer method. *Journal of Chemical Information and Modeling*. 2022;62(2):309-23.
- [188] Fajer M, Hamelberg D, McCammon JA. Replica-exchange accelerated molecular dynamics (REXAMD) applied to thermodynamic integration. *Journal of chemical theory and computation*. 2008;4(10):1565-9.
- [189] Jiang W, Roux B. Free energy perturbation Hamiltonian replica-exchange molecular dynamics (FEP/H-REMD) for absolute ligand binding free energy calculations. *Journal of chemical theory and computation*. 2010;6(9):2559-65.

- [190] Meli M, Colombo G. A Hamiltonian replica exchange molecular dynamics (MD) method for the study of folding, based on the analysis of the stabilization determinants of proteins. *International journal of molecular sciences*. 2013;14(6):12157-69.
- [191] Martire S, Decherchi S, Cavalli A. OBIWAN: An Element-Wise Scalable Feed-Forward Neural Network Potential. *Journal of Chemical Theory and Computation*. 2024;20(14):6287-302.
- [192] Werner P, Hartmann AK. Similarity of extremely rare nonequilibrium processes to equilibrium processes. *Physical Review E*. 2021;104(3):034407.
- [193] Híjar H, Quintana-H J, Sutmann G. Non-equilibrium work theorems for the two-dimensional Ising model. *Journal of statistical mechanics: theory and experiment*. 2007;2007(04):P04010.
- [194] Hartmann AK. High-precision work distributions for extreme nonequilibrium processes in large systems. *Physical Review E*. 2014;89(5):052103.
- [195] Chatelain C, Karevski D. Probability distributions of the work in the two-dimensional Ising model. *Journal of Statistical Mechanics: Theory and Experiment*. 2006;2006(06):P06005.
- [196] Spitaleri A, Decherchi S, Cavalli A, Rocchia W. Fast dynamic docking guided by adaptive electrostatic bias: The MD-binding approach. *Journal of chemical theory and computation*. 2018;14(3):1727-36.
- [197] Zia SR, Gaspari R, Decherchi S, Rocchia W. Probing hydration patterns in class-A GPCRs via biased MD: the A2A receptor. *Journal of Chemical Theory and Computation*. 2016;12(12):6049-61.
- [198] Gobbo D, Piretti V, Di Martino RMC, Tripathi SK, Giabbai B, Storici P, et al. Investigating drug–target residence time in kinases through enhanced sampling simulations. *Journal of Chemical Theory and Computation*. 2019;15(8):4646-59.

- [199] Ferrarotti MJ, Rocchia W, Decherchi S. Finding principal paths in data space. *IEEE transactions on neural networks and learning systems*. 2018;30(8):2449-62.
- [200] Doudou S, Burton NA, Henchman RH. Standard free energy of binding from a one-dimensional potential of mean force. *Journal of chemical theory and computation*. 2009;5(4):909-18.
- [201] Decherchi S, Rocchia W. A general and Robust Ray-Casting-Based Algorithm for Triangulating Surfaces at the Nanoscale. *Plos One*. 2013 04;8(4):1-15. Available from: <https://doi.org/10.1371/journal.pone.0059744>.
- [202] Decherchi S, Bottegoni G, Spitaleri A, Rocchia W, Cavalli A. BiKi life sciences: a new suite for molecular dynamics and related methods in drug discovery. *Journal of chemical information and modeling*. 2018;58(2):219-24.
- [203] Efron B. Bootstrap Methods: Another Look at the Jackknife. In: Kotz S, Johnson NL, editors. *Breakthroughs in Statistics: Methodology and Distribution*. New York, NY: Springer New York; 1992. p. 569-93.
- [204] Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. 2015;1:19-25.
- [205] Tribello GA, Bonomi M, Branduardi D, Camilloni C, Bussi G. PLUMED 2: New feathers for an old bird. *Computer physics communications*. 2014;185(2):604-13.
- [206] Essex JW, Severance DL, Tirado-Rives J, Jorgensen WL. Monte Carlo simulations for proteins: Binding affinities for trypsin- benzamidine complexes via free-energy perturbations. *The Journal of Physical Chemistry B*. 1997;101(46):9663-9.
- [207] Mares-Guia M, Shaw E. Studies on the active center of trypsin: the binding of amidines and guanidines as models of the substrate side chain. *Journal of biological chemistry*. 1965;240(4):1579-85.

- [208] Buch I, Giorgino T, De Fabritiis G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences*. 2011;108(25):10184-9.
- [209] Seeliger MA, Nagar B, Frank F, Cao X, Henderson MN, Kuriyan J. c-Src binds to the cancer drug imatinib with an inactive Abl/c-Kit conformation and a distributed thermodynamic penalty. *Structure*. 2007;15(3):299-311.
- [210] Patel JS, Berteotti A, Ronsisvalle S, Rocchia W, Cavalli A. Steered Molecular Dynamics Simulations for Studying Protein-Ligand Interaction in Cyclin-Dependent Kinase 5. *Journal of Chemical Information and Modeling*. 2014 Feb;54(2):470-80. Available from: <https://doi.org/10.1021/ci4003574>.
- [211] Fu H, Chen H, Blazhynska M, Goulard Coderc de Lacam E, Szczepaniak F, Pavlova A, et al. Accurate determination of protein: ligand standard binding free energies from molecular dynamics simulations. *Nature protocols*. 2022;17(4):1114-41.
- [212] Berg S, Bergh M, Hellberg S, Högdin K, Lo-Alfredsson Y, Söderman P, et al. Discovery of novel potent and highly selective glycogen synthase kinase-3 β (GSK3 β) inhibitors for Alzheimer's disease: design, synthesis, and characterization of pyrazines. *Journal of medicinal chemistry*. 2012;55(21):9107-19.
- [213] Mobley DL, Gilson MK. Predicting binding free energies: frontiers and benchmarks. *Annual review of biophysics*. 2017;46(1):531-58.
- [214] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*. 2018;46(W1):W296-303.
- [215] Febrer Martinez P, Rizzi V, Aureli S, Gervasio FL. Host-Guest binding free energies à la carte: an automated OneOPES protocol. *bioRxiv*. 2024:2024-08.
- [216] Capelli R, Carloni P, Parrinello M. Exhaustive search of ligand binding pathways via volume-based metadynamics. *The journal of physical chemistry letters*. 2019;10(12):3495-9.

- [217] Rydzewski J, Valsson O. Finding multiple reaction pathways of ligand unbinding. *The Journal of Chemical Physics*. 2019;150(22).
- [218] Boubeta FM, Contestín García RM, Lorenzo EN, Boechi L, Estrin D, Sued M, et al. Lessons learned about steered molecular dynamics simulations and free energy calculations. *Chemical biology & drug design*. 2019;93(6):1129-38.
- [219] Xie T, Saleh T, Rossi P, Kalodimos CG. Conformational states dynamically populated by a kinase determine its function. *Science*. 2020;370(6513):eabc2754.
- [220] Nolen B, Taylor S, Ghosh G. Regulation of protein kinases: controlling activity through activation segment conformation. *Molecular cell*. 2004;15(5):661-75.
- [221] Ayaz P, Lyczek A, Paung Y, Mingione VR, Iacob RE, de Waal PW, et al. Structural mechanism of a drug-binding process involving a large conformational change of the protein target. *Nature communications*. 2023;14(1):1885.
- [222] Warner KD, Hajdin CE, Weeks KM. Principles for targeting RNA with drug-like small molecules. *Nature reviews Drug discovery*. 2018;17(8):547-58.
- [223] Rizvi NF, Smith GF. RNA as a small molecule druggable target. *Bioorganic & Medicinal Chemistry Letters*. 2017;27(23):5083-8.
- [224] Falese JP, Donlic A, Hargrove AE. Targeting RNA with small molecules: from fundamental principles towards the clinic. *Chemical Society Reviews*. 2021;50(4):2224-43.
- [225] Di Giorgio A, Duca M. Synthetic small-molecule RNA ligands: future prospects as therapeutic agents. *MedChemComm*. 2019;10(8):1242-55.
- [226] Connelly CM, Moon MH, Schneekloth JS. The emerging role of RNA as a therapeutic target for small molecules. *Cell chemical biology*. 2016;23(9):1077-90.

- [227] Tong Y, Childs-Disney JL, Disney MD. Targeting RNA with small molecules, from RNA structures to precision medicines: IUPHAR review: 40. *British Journal of Pharmacology*. 2024;181(21):4152-73.
- [228] Dhillon S. Risdiplam: first approval. *Drugs*. 2020;80:1853-8.
- [229] Sheridan C. First small-molecule drug targeting RNA gains momentum. *Nature biotechnology*. 2021;39(1):6-9.
- [230] Blakeley BD, DePorter SM, Mohan U, Burai R, Tolbert BS, McNaughton BR. Methods for identifying and characterizing interactions involving RNA. *Tetrahedron*. 2012;68(43):8837-55.
- [231] Manigrasso J, Marcia M, De Vivo M. Computer-aided design of RNA-targeted small molecules: a growing need in drug discovery. *Chem*. 2021;7(11):2965-88.
- [232] Bernetti M, Aguti R, Bosio S, Recanatini M, Masetti M, Cavalli A. Computational drug discovery under RNA times. *QRB discovery*. 2022;3:e22.
- [233] Kavita K, Breaker RR. Discovering riboswitches: the past and the future. *Trends in biochemical sciences*. 2023;48(2):119-41.
- [234] Ellinger E, Chauvier A, Romero RA, Liu Y, Ray S, Walter NG. Riboswitches as therapeutic targets: promise of a new era of antibiotics. *Expert Opinion on Therapeutic Targets*. 2023;27(6):433-45.
- [235] Panchal V, Brenk R. Riboswitches as drug targets for antibiotics. *Antibiotics*. 2021;10(1):45.
- [236] Kallert E, Fischer TR, Schneider S, Grimm M, Helm M, Kersten C. Protein-based virtual screening tools applied for RNA–ligand docking identify new binders of the PreQ1-riboswitch. *Journal of Chemical Information and Modeling*. 2022;62(17):4134-48.
- [237] Howe JA, Wang H, Fischmann TO, Balibar CJ, Xiao L, Galgoci AM, et al. Selective small-molecule inhibition of an RNA structural element. *Nature*. 2015;526(7575):672-7.

- [238] Wang Y, Parmar S, Schneekloth JS, Tiwary P. Interrogating RNA–small molecule interactions with structure probing and artificial intelligence-augmented molecular simulations. *ACS Central Science*. 2022;8(6):741-8.
- [239] Connelly CM, Numata T, Boer RE, Moon MH, Sinniah RS, Barchi JJ, et al. Synthetic ligands for PreQ1 riboswitches provide structural and mechanistic insights into targeting RNA tertiary structure. *Nature communications*. 2019;10(1):1501.
- [240] Bešker N, Gervasio FL. Using metadynamics and path collective variables to study ligand binding and induced conformational transitions. *Computational drug discovery and design*. 2012:501-13.
- [241] Provasi D, Artacho MC, Negri A, Mobarec JC, Filizola M. Ligand-induced modulation of the free-energy landscape of G protein-coupled receptors explored by adaptive biasing techniques. *PLoS computational biology*. 2011;7(10):e1002193.
- [242] Callea L, Bonati L, Motta S. Metadynamics-based approaches for modeling the hypoxia-inducible factor 2α ligand binding process. *Journal of Chemical Theory and Computation*. 2021;17(7):3841-51.
- [243] Lin YL, Meng Y, Huang L, Roux B. Computational study of Gleevec and G6G reveals molecular determinants of kinase inhibitor selectivity. *Journal of the American Chemical Society*. 2014;136(42):14753-62.
- [244] Piana S, Donchev AG, Robustelli P, Shaw DE. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *The journal of physical chemistry B*. 2015;119(16):5113-23.
- [245] Tan D, Piana S, Dirks RM, Shaw DE. RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proceedings of the National Academy of Sciences*. 2018;115(7):E1346-55.
- [246] Smith JS, Nebgen B, Lubbers N, Isayev O, Roitberg AE. Less is more: Sampling chemical space with active learning. *The Journal of chemical physics*. 2018;148(24).

- [247] Seeliger MA, Ranjitkar P, Kasap C, Shan Y, Shaw DE, Shah NP, et al. Equally potent inhibition of c-Src and Abl by compounds that recognize inactive kinase conformations. *Cancer research*. 2009;69(6):2384-92.
- [248] Agafonov RV, Wilson C, Otten R, Buosi V, Kern D. Energetic dissection of Gleevec’s selectivity toward human tyrosine kinases. *Nature structural & molecular biology*. 2014;21(10):848-53.
- [249] Shan Y, Seeliger MA, Eastwood MP, Frank F, Xu H, Jensen MØ, et al. A conserved protonation-dependent switch controls drug binding in the Abl kinase. *Proceedings of the National Academy of Sciences*. 2009;106(1):139-44.
- [250] Schindler T, Bornmann W, Pellicena P, Miller WT, Clarkson B, Kuriyan J. Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. *Science*. 2000;289(5486):1938-42.
- [251] Lovera S, Morando M, Pucheta-Martinez E, Martinez-Torrecuadrada JL, Saladino G, Gervasio FL. Towards a molecular understanding of the link between imatinib resistance and kinase conformational dynamics. *PLoS computational biology*. 2015;11(11):e1004578.
- [252] Ganser LR, Kelly ML, Herschlag D, Al-Hashimi HM. The roles of structural dynamics in the cellular functions of RNAs. *Nature reviews Molecular cell biology*. 2019;20(8):474-89.
- [253] Bosio S, Bernetti M, Rocchia W, Masetti M. Similarities and Differences in Ligand Binding to Protein and RNA Targets: The Case of Riboflavin. *Journal of Chemical Information and Modeling*. 2024.
- [254] Rasouli A, Pickard IV FC, Sur S, Grossfield A, Isik Bennett M. Essential Considerations for Free Energy Calculations of RNA–Small Molecule Complexes: Lessons from the Theophylline-Binding RNA Aptamer. *Journal of Chemical Information and Modeling*. 2024;65(1):223-39.
- [255] Štoček JR, Dračinský M. Tautomerism of guanine analogues. *Biomolecules*. 2020;10(2):170.

- [256] Bhati AP, Wan S, Coveney PV. Equilibrium and Nonequilibrium Ensemble Methods for Accurate, Precise and Reproducible Absolute Binding Free Energy Calculations. *Journal of Chemical Theory and Computation*. 2024;21(1):440-62.
- [257] Krasowski A, Muthas D, Sarkar A, Schmitt S, Brenk R. DrugPred: a structure-based approach to predict protein druggability developed using an extensive nonredundant data set. *Journal of chemical information and modeling*. 2011;51(11):2829-42.
- [258] Nicolaou K. Advancing the drug discovery and development process. *Angewandte Chemie*. 2014;126(35):9280-92.
- [259] Aguti R, Gardini E, Bertazzo M, Decherchi S, Cavalli A. Probabilistic pocket druggability prediction via one-class learning. *Frontiers in Pharmacology*. 2022;13:870479.
- [260] Decherchi S, Rocchia W. A general and robust ray-casting-based algorithm for triangulating surfaces at the nanoscale. *PloS one*. 2013;8(4):e59744.
- [261] Phillips M, Georgiev I, Dehof AK, Nickels S, Marsalek L, Lenhof HP, et al. Measuring properties of molecular surfaces using ray casting. In: 2010 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW). IEEE; 2010. p. 1-7.
- [262] Edelsbrunner H. Deformable smooth surface design. *Discrete & Computational Geometry*. 1999;21:87-115.
- [263] Zhang Y, Xu G, Bajaj C. Quality meshing of implicit solvation models of biomolecular structures. *Computer aided geometric design*. 2006;23(6):510-30.
- [264] Bakan A, Meireles LM, Bahar I. ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics*. 2011;27(11):1575-7.
- [265] Zhang S, Krieger JM, Zhang Y, Kaya C, Kaynak B, Mikulska-Ruminska K, et al. ProDy 2.0: increased scale and scope after 10 years of protein dynamics modelling with Python. *Bioinformatics*. 2021;37(20):3657-9.

- [266] Rego N, Koes D. 3Dmol. js: molecular visualization with WebGL. *Bioinformatics*. 2015;31(8):1322-4.
- [267] Mollica L, Decherchi S, Zia SR, Gaspari R, Cavalli A, Rocchia W. Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations. *Scientific reports*. 2015;5(1):11539.
- [268] Mollica L, Theret I, Antoine M, Perron-Sierra F, Charton Y, Fourquez JM, et al. Molecular dynamics simulations and kinetic measurements to estimate and predict protein–ligand residence times. *Journal of medicinal chemistry*. 2016;59(15):7167-76.

Ringraziamenti

Chi l'avrebbe mai detto che saremmo arrivati fin qui? In questo posto, con le mie persone, a festeggiare insieme il traguardo del *Dottorato*. La Ele, il Gian e tutti noi del passato non l'avremmo mai potuto immaginare. È incredibile pensare che da una scelta fatta a caso, senza nemmeno sapere davvero cosa fosse la *Chimica*, mi ritrovi oggi qui a scrivere i ringraziamenti della mia tesi di *Dottorato*. Incredibile, ma bellissimo.

Per questo, arrivati insieme a questa giornata, voglio ringraziarvi. Inanzitutto, se stai leggendo queste parole significa che, in un modo o nell'altro, sei stato parte di questo lungo e strano viaggio: quindi **Grazie**. Grazie per avermi sopportata e supportata lungo il cammino, per aver sicuramente raccolto qualche pezzettino quando serviva, asciugato qualche lacrima e festeggiato con me anche un piccolo traguardo.

Scendendo più nel dettaglio, il mio primo grazie va al Professor Cavalli, a Sergio Decherchi e ad Alessia Ghidini, per avermi dato l'opportunità di lavorare e crescere professionalmente accanto a voi.

Un ringraziamento speciale va al mio compagno di avventure, al mio “fratello discombobulated”, con cui ho condiviso scelte opinabili e momenti indimenticabili. Senza di te questo percorso non sarebbe stato lo stesso, se non impossibile. Grazie per la tua leggerezza e profondità. Come ben sai, sei un fratello e il “capitano secondo” del buon Gianni. Un cacciatore incredibile di girassoli. La Ele e il Gian del presente ringraziano la Ele e il Gian del passato, per essersi scelti, per aver camminato insieme e per aver costruito, la Ele e il Gian del futuro.

In secondo luogo, ma non certo per importanza, grazie alla mia famiglia; pilastro fondamentale di questo lungo viaggio che è la vita. A mia mamma e a mia sorella: siete sempre state un sostegno imprescindibile, il mio punto fermo, e sarete per sempre la mia luce di riferimento in mezzo al mare.

Mamma *Effe*, anche se a volte dura e testarda, sei riuscita a renderci ciò che siamo e a permetterci di diventare ciò che saremo. Per questo, ti dico il mio primo grande **Grazie**. Ci hai trasmesso valori, convinzioni e pensieri che spesso non si possono spiegare a parole. Ci hai insegnato tanto e, anche se a volte non lo ammetto, sarai sempre (ahimè, purtroppo) la mia più grande fonte d'ispirazione. Anche in questo tuo ruolo così difficile sei riuscita a fare "mille cose insieme": sei una mamma, un'amica, una compagna di avventure (mi dispiace non di bevute!) e una donna straordinaria. Per tutto questo, ti dico un altro grande grazie dal profondo del mio cuore.

Marta, senza di te tutto sarebbe stato diverso. Anche se può sembrare scontato o banale, il mio primo grazie nei tuoi confronti è un grazie infinito per esserci. Lo urlerei a squarciagola, solo per far sì che tutto il mondo si fermi ad ascoltarlo e per poterci guardare insieme. Grazie per la tua calma, la tua saggezza e il tuo equilibrio. Grazie per aver creduto in me anche quando io non ci riuscivo più. Grazie per aver saputo vedere il bene nei miei gesti, anche quando era nascosto e offuscato dalle mie mille follie. Grazie per la pazienza con cui cerchi di capire la mia testa un po' matta, per l'impegno che metti ogni giorno nel farci crescere insieme e per la naturalezza con cui hai scelto di far parte del mio mondo così diverso, a volte, dal tuo. Attraverso i tuoi occhi felici e grati alla vita, a volte riesco ancora a intravedere il sorriso di papà. In questo mare in tempesta, tu sei la barchetta che mi riporta sempre a riva.

Poi, grazie a Francesco. Grazie per la pazienza che trovi ogni giorno, per il tuo supporto costante e per la fiducia che non mi hai mai fatto mancare. Grazie perché, ad oggi, hai condiviso con me ogni tassello importante di questa imprevedibile vita. Tra sorrisi, pianti, urla e paranoie, grazie per essere rimasto accanto a me anche nei momenti più difficili. *Più di ieri, meno di domani.*

Poi, grazie a Giorgia, che con la tua schiettezza, il tuo cinismo e la tua profonda sincerità mi hai sempre aiutata a guardare il mondo in modo chiaro e spesso più leggero. Non c'è momento di felicità o di sconforto che vorrei vivere senza di te. Così diverse, eppure così profondamente unite. Se ci penso, mi luccicano gli occhi dall'emozione. Grazie per il tuo credere in me costante, per essere la mia più grande sostenitrice. Spero che tu possa vedere per sempre la stessa figura in me. La nostra amicizia è una di quelle rare, che non hanno bisogno di parole per capirsi, ma solo di uno sguardo per sentirsi a casa.

Accanto a te, ringrazio anche le mie ragazze del cuore: Marta, Valeria e Mara. Grazie per essere semplicemente voi, *Bottanazze* del mio cuore, e per tutto ciò che siamo insieme. Grazie per la spensieratezza, l'amore e il calore che non mi avete mai fatto mancare. Ognuna di voi ha un posto speciale nel mio cuoricino, e non avrei potuto desiderare di riservarlo a nient' altro. Grazie, quindi, per esservi aggiudicate questo piccolo spazio di mio cuore.

In aggiunta a questo, però, voglio dedicare un grazie un po' diverso: grazie a Marta. Grazie perché ascolti, comprendi e provi, con pazienza e affetto, a dare un senso a ogni mia follia. Non è facile, per una folle come me, sentirsi a casa e non giudicata da occhi esterni dai suoi. Per questo, ti dedico il mio più sincero e profondo grazie. Grazie perché cerchi di vedere il mondo con i miei occhi e, nonostante tu abbia conosciuto anche le parti mie più oscure, continui a scegliermi ogni giorno accanto a te. Grazie perché in te ho trovato molto più di un'amica: una sorella, una spalla, e soprattutto un'incredibile compagna di avventure.

Non può mancare poi un grazie a Chiara e Alice, che con la loro calma, amicizia sincera e profondità mi hanno accompagnata lungo questo percorso con costante affetto. Quest'anno ho realizzato che siete con me da metà della mia vita, e sono profondamente grata che abbiate colorato e condiviso con me ogni passo di questo cammino. Siamo cresciute, cambiate e diventate grandi insieme (Chiara forse un po' più di noi!), e spero con tutto il cuore che il futuro ci riservi ancora tanti traguardi e esperienze insieme. In particolare, Chiara, grazie di cuore per tutto ciò che abbiamo creato e che stiamo continuando a costruire insieme. Averti accanto è un dono per cui sono immensamente grata.

Poi, grazie a Carlo, perché sei stato fondamentale. Grazie per aver remato con me, per aver capito, per aver creduto e per avermi spinto a non mollare. Condividiamo, ahimè, una parte buia dell'anima, ma insieme a te mi è sembrata meno cupa e più affrontabile. Sei, senza dubbio, il dono più grande di questo dottorato.

E ancora, grazie a Bea, Gaia, Cami, Ale, Betto, Frullo, Luca, Ila, Campa e a tutti voi che avete colorato la mia vita e reso questo cammino più luminoso. Con voi ogni giorno è stato migliore. Un grazie speciale a Vala, Gabri e Nevio: grazie per le risate che mi avete regalato, per la gioia che mi avete trasmesso e per la bellezza che riuscite a trovare (e a creare) in ogni cosa. Con voi, anche il mondo più complicato sembra sempre più semplice, più autentico, più bello. Vivir para compartir, fratelli miei. La felicità, con voi, ha sempre avuto il sapore di casa.

Quindi, alla mia squadra formidabile di amici, **Grazie**. Grazie perché, oggi più che mai, penso che non avrei potuto chiedere o desiderare di meglio dalla vita.

Infine, concludo tutto questo con una frase che amo da morire (e molti di voi già lo sanno): *Gli amici che ti fai quando hai quindici, diciassette, vent'anni hanno una qualità, una specialità, che nella vita poi sarà irripetibile. Avrai altre amicizie, anche molto qualificate, ma qualcuno che ti sia stato testimone quando potevi ancora essere tutto, quello non si ripete.*

Beh ecco, penso che tutti voi, amici miei che state leggendo queste parole, in un modo o nell'altro, siete stati testimoni di una parte di me. E con il vostro piccolo o grande contributo avete reso possibile il cammino che mi ha portata ad oggi fino a qui. Quindi, a voi rivolgo la mia più profonda gratitudine.

Infine, grazie a te, saggio *capitano*. Mi hai ispirata e sostenuta come nessun altro potrà mai più fare. Crescendo, mi sorprendo ogni giorno nel riconoscere quante piccole parti di te vivano ancora dentro di me. Grazie, perché ciò che hai trasmesso e lasciato è qualcosa di raro: il bene autentico, l'amore sincero, il modo di sentire profondamente le cose e la voglia di cambiare (anche solo un po') questo mondo così complesso. A volte spero che tu possa vedere, o almeno immaginare, quanto cammino stiamo percorrendo, senza mai dimenticare di rivolgere un pensiero e un bellissimo brindisi a te.

Nelle giornate di tempesta, cerco ancora il tuo conforto.

bw: