

Politecnico di Torino

Politecnico di Torino

Master's Degree in Physics of Complex Systems

Department of Applied Science and Technology

Advanced Molecular Dynamics Simulations of Cyclic Peptide-Protein Complexes

Supervisors
Prof. Dirk Stratmann
Prof. Fabio Pietrucci
Dr. Valerio Rizzi
Prof. Carla Bosia

Candidate
Alessandro Iannitti

Academic Year 2025/2026

Abstract

Cyclic peptides are an important class of pharmaceutical drugs, offering an optimal balance of high affinity, specificity and metabolic stability. Their computational modeling remains, however, challenging due to the complex interplay between conformational flexibility and structural rigidity. In this thesis we investigate the binding-unbinding mechanisms of cyclic peptide-protein complexes to accurately predict their absolute binding free energies (ΔG_b°) using advanced molecular dynamics (MD) simulations. We focused primarily on the Kelch domain of human KEAP1 bound to different Nrf2-based cyclic peptides (these complexes are called 7K2*) and then validated our approach on a highly rigid and larger system (1SFI). In order to overcome the sampling limitations of classical MD, we combined a Hamiltonian Replica Exchange (HREX) setup driven by On-the-fly Probability Enhanced Sampling (OneOPES) with Funnel Metadynamics, efficiently biasing hydration collective variables, while restraining peptide exploration in the unbound state. We have tested two force fields, Amber14SB and a99SB-disp, for each complex, revealing a profound structural bias: classical force fields like Amber14SB work well on rigid systems but over-compact flexible peptides in the unbound state, thus overestimating the binding free energy. On the other hand, a99SB-disp accurately models flexible peptides but fails on rigid ones. To overcome this limitation we developed a predictive three-level screening protocol based on explicit and implicit-solvent REMD simulations to quantify the isolated peptide flexibility. Through a Kernel Density Estimation (KDE) analysis, this protocol directly assigns the whole complex to a specific force field or a hybrid consensus approach, yielding excellent agreement with experimental data. Additionally, we establish a rigorous convergence assessment protocol, by coupling the stabilization of the absolute binding free energy ($\Delta G_b^\circ(t)$) with the RMSD of the free energy surface (FES-RMSD) and sampling efficiency metrics. Finally, we present preliminary results of an advanced protocol aimed at overcoming the binary choice of force fields by directly biasing internal conformational collective variables of the peptide in the bulk solvent to prevent over-compaction in the unbound state.

Contents

Abstract **i**

1 Introduction **1**
1.1 The Role of Cyclic Peptides in Drug Discovery 1
1.2 Computational Estimation of Binding Free Energies 2
1.2.1 Rigorous Methods vs. Approximate Approaches 3
1.2.2 The Force Field Dichotomy: Over-compaction vs. Under-binding 5
1.3 Aims and Structure of the Thesis 6

2 Methods **8**
2.1 Computational Techniques and Free Energy Engines 8
2.1.1 Molecular Dynamics, GROMACS and Force Fields Overview . . 8
2.1.2 OneOPES: On-the-fly Probability Enhanced Sampling 10
2.1.3 Funnel Metadynamics for Absolute Binding Free Energy 12
2.2 Biological Systems 13
2.2.1 The Target Dataset: 7K2* Protein-Peptide Complexes 14
2.2.2 First Validation System: 1SFI 16
2.3 Multi-Replica Simulation Setup 17
2.3.1 The Fingerprint Approach for Solvation 18
2.4 Thermodynamic Analysis Protocol 20
2.4.1 Multi-Run Convergence: ΔG_b° Integration and Cross-Replica Consistency 21
2.4.2 Single-Run Convergence: FES-RMSD and Sampling Efficiency . 22

3 Results and Discussion **24**
3.1 The Force Field Dichotomy: Baseline Performance 24
3.1.1 Amber14SB: Over-compaction and Over-binding 25
3.1.2 a99SB-disp: Under-binding for Rigid Complexes 29
3.2 A Predictive Three-Level Screening Protocol 33
3.2.1 Conformational Sampling in Explicit and Implicit Solvent 34
3.2.2 Protocol details 35
3.3 Application of the Predictive Screening Protocol 39
3.4 1SFI: an Out-of-Family Rigid Complex 42
3.5 Beyond Heuristics: Preliminary Results of the Advanced Protocol 46

4 Conclusions and Future Perspectives **52**

A Computational Techniques **55**
A.1 PLUMED: The Core Engine 55
A.2 Hamiltonian Replica Exchange (HREX) Architecture 56

B Conformational sampling **57**

C Kernel Density Estimation **61**

D Leave-One-Out Cross Validation of the Screening Protocol **63**

E Screening Protocol with Alternative Flexibility Metrics **64**

F Computational details **66**

G Simulations Results **68**
G.1 Results with Amber14SB 68
G.2 Results with a99SB-disp 72
G.3 Application of the Screening Protocol 73

Bibliography **75**

Chapter 1

Introduction

1.1 The Role of Cyclic Peptides in Drug Discovery

Proteins provide the fundamental framework for all biological processes of living organisms as they are heavily involved in cell function, including structural support, catalysis, signal transduction and molecule transport [1]. However, since the crowded cellular environments limit their action to short-range interactions, cells rely on an elaborate network of contacts between proteins called interactome, which allows them to communicate and coordinate across the cell: this makes them capable of performing the complex functions required for life [2, 3]. The interactions between two or more proteins within these networks are called Protein-Protein Interactions (PPIs). PPIs play a central role in numerous biological processes, including signal transduction, protein function regulation and, most importantly, disease development, as they govern viral diffusion and virus survival within host cells [4]. Mediating these interactions with selected PPI mediators is particularly important to solve many clinical conditions caused by PPIs [1]. Cyclic peptides have gathered attention over the past decades as an important class of pharmaceutical drugs and as an alternative to small molecules, as they interfere with PPIs particularly well [4–9]. Overall, we can divide PPIs mediators in three categories:

- **Small molecules:** These are low molecular weight organic compounds designed to prompt a specific biological process within the body and account for over 90% of marketed drugs [10]. They possess relatively simple structures, are generally stable, and their predictable in vivo behaviour allows for straightforward, often oral, dosing protocols. They exert their therapeutic effects in many ways: they can act as enzyme inhibitors, to block the activity of proteins and can interfere with disease processes (an example is provided by statins which inhibit the activity of an enzyme involved in the production of cholesterol in the liver), they can work as receptor agonist/antagonists or as ion channel modulators [10]. They can also modulate PPIs: for instance, venetoclax is an approved PPI drug used for acute myeloid leukemia (AML) and chronic lymphocytic leukemia (CLL) that works by disrupting the interaction between two proteins that would otherwise work together to promote the survival of cancer cells [11].

- **Cyclic peptides:** These are polypeptide chains composed of canonical and non-canonical amino acids that are connected at distant positions to form macrocyclic structures [9]. In drug development, cyclic peptides occupy a middle space between small molecules and biologics, in fact they have diverse shapes, sizes and

chemical compositions which allow them to bind several targets, including PPIs, that often lack defined pockets for small molecule binding and are instead characterized by flat, featureless surfaces [12]. They typically mediate PPIs through structural mimicry: since protein partners often interact through specific secondary structural motifs (α-helices, β-strands), cyclic peptides can be designed to mimic these interactions as shown by Robinson and co-workers [13]. Moreover, cyclic peptides bind to targets with higher affinity and specificity than their linear analogs [14], also because their constrained structure reduces the entropic penalty incurred upon binding. Finally, cyclization provides enhanced protection against enzymatic degradation compared to linear peptides [9, 15].

- **Biologics:** These represent a diverse and highly complex class of therapeutics that are produced by living organisms. There are three major categories of biologics: endogenous proteins (like insulin), Fc-fusion proteins and monoclonal antibodies (mAbs) [16]. Because of their large size ($>$ 140 kDa for antibodies) and complex three-dimensional structures, they can engage targets with high specificity and affinity, often binding to extended PPI surfaces that are considered *undruggable* by small molecules [17, 18].

When comparing these three groups, cyclic peptides emerge as a great candidate in modern drug discovery. They occupy a therapeutic *middle space* that combines the most useful characteristics of both small molecules and biologics. Small molecules have been historically unable to target flat, featureless PPI interfaces, leaving a vast portion of the human interactome *undruggable* [10], even if recently new methods are emerging to try to bridge these gaps [19]. Biologics, on the other hand, possess the necessary size and complexity to drug those interfaces but they are burdened by complex manufacturing processes, severe immunogenic risks and a strict confinement to extracellular targets due to their inability to cross cell membranes [17]. Cyclic peptides address these problems: like biologics, they bind extended PPI surfaces with high affinity and specificity. Yet, much like small molecules, cyclic peptides can be efficiently produced via chemical synthesis. Furthermore, with careful physicochemical optimization, such as reducing the polar surface area or utilizing N-methylated amino acids (as done in [20]), cyclic peptides can successfully cross cellular membranes, unlocking the ability to drug intracellular PPIs [9].

1.2 Computational Estimation of Binding Free Energies

Binding affinity is a crucial metric in drug discovery, representing the strength of the interaction between a ligand and its biological target. While binding affinity can be determined with several experimental techniques, including biophysical methods such as surface plasmon resonance (SPR) and isothermal titration calorimetry (ITC), as well as biochemical methods like radioactive and enzymatic assays, they are inherently limited [21]. Experimental screening of vast compound libraries requires synthesis and purification of each molecule; this process is rather time and cost-demanding [22]. Moreover, even though experiments provide macroscopic information on the thermodynamic parameters, they usually lack detailed atomistic description that helps to rationally optimize a lead compound [21, 22]. This issue becomes especially important when designing cyclic peptides that aim to disrupt PPIs. The design of such macrocycles involves precise regulation of their conformational ensemble in order to minimize the entropic contribution upon binding while maximizing the enthalpic contribution to

the binding across wide, flat protein interfaces [9]. To address the limitations in experimental techniques, computational methods have become increasingly popular. The binding affinity (K_a) is directly related to the standard binding free energy (ΔG_b°), which measures the difference in free energy between the bound and unbound states through the following equation:

$$\Delta G_b^\circ = -\frac{1}{\beta} \ln (K_a C^\circ) \tag{1.1}$$

where $\beta = (k_B T)^{-1}$ and $C^\circ = 1$ mol/L is the standard-state concentration. Accurately estimating ΔG_b° allows researchers to rank potential drug candidates prior to synthesis and to gain information about the binding mechanisms, thus also accelerating the discovery pipeline [21].

1.2.1 Rigorous Methods vs. Approximate Approaches

Estimating binding free energies computationally involves a constant trade-off between accuracy and computational cost. Fast, approximate methods, such as molecular docking combined with end-point scoring functions like MM/PBSA (Molecular Mechanics/ Poisson Boltzmann Surface Area) [23], are frequently employed to rapidly screen large libraries. While these approaches have shown promising results in predicting protein-peptide interactions and their binding affinities [24], their accuracy is affected by approximations in solvation modeling and entropy estimation [25]; that's why performance of MM/PBSA methods differs across various systems [26, 27]. For flexible molecules like cyclic peptides, neglecting the entropic cost given by the restriction of conformational space upon binding can lead to inaccurate results in predicting binding affinities [22]. Since binding and unbinding are rare events on standard molecular dynamics (MD) timescales, rigorous methods rely on enhanced sampling techniques, which fall into two categories: alchemical transformations and path-based methods [21].

Alchemical transformations. Alchemical methods, such as Free Energy Perturbation (FEP) and Thermodynamic Integration (TI), rely on a hybrid Hamiltonian obtained by interpolating the potential energy from an initial state (A) to a final state (B), through a non-physical coupling parameter λ:

$$V(\mathbf{q}; \lambda) = (1 - \lambda) V_A(\mathbf{q}) + \lambda V_B(\mathbf{q}) \tag{1.2}$$

where $0 \leq \lambda \leq 1$. The system is then sampled at several values of λ and finally the binding free energy is obtained using TI or FEP:

$$\Delta G_{AB}^{TI} = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial V_\lambda}{\partial \lambda} \right\rangle_\lambda d\lambda \qquad \Delta G_{AB}^{FEP} = -\frac{1}{\beta} \ln \langle \exp(-\beta \Delta V_{AB}) \rangle_\lambda^{\text{eq}} \tag{1.3}$$

In practice, to compute absolute binding free energies, these methods utilize a "double decoupling" thermodynamic cycle (figure 1.1). In this cycle, the ligand is restrained in the bulk solvent, then decoupled (interactions are switched off, λ goes from 1 to 0) from the bulk solvent and the protein ($L \rightarrow L^\star \rightarrow 0$). Subsequently the ligand is moved in the bound state, re-coupled (interactions are switched on, λ goes from 0 to 1) and finally we remove the restraints. Then, the absolute binding free energy is the sum of the ΔG at each step. Alchemical methods are the standard to compute relative binding free energies ($\Delta\Delta G$) but they have some important limitations: they require significant phase-space overlap between intermediate λ-states; thus, decoupling a large,

flexible macrocycle with high desolvation penalties can lead to severe convergence issues. Moreover, since we are dealing with non-physical thermodynamic paths, alchemical transformations lack the ability to provide kinetic or mechanistic insights into the binding process.

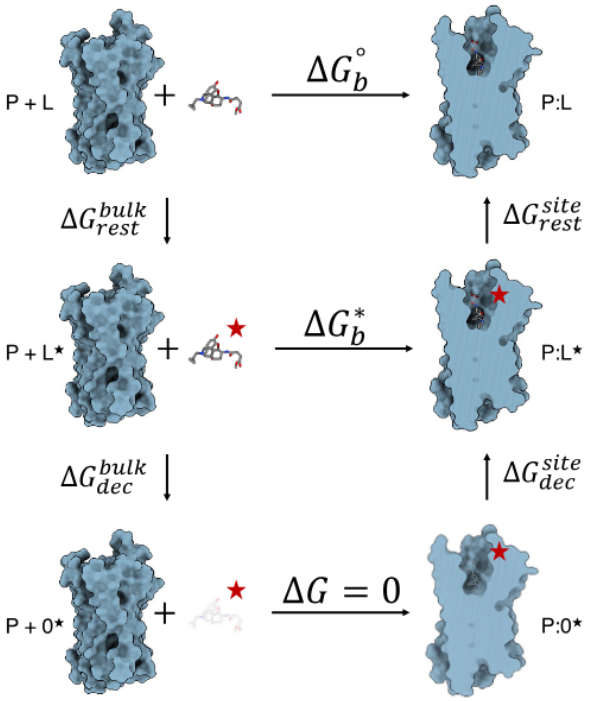


Figure 1.1: Schematic representation of a thermodynamic cycle using double decoupling for absolute binding free energy calculations. The ligand (L) is restrained in its bound conformation, position and orientation. The restrained ligand (L*) is decoupled from both the protein and bulk solvent (denoted by 0). Then the ligand is moved to the bound state, it is recoupled and restraints are removed. Following the thermodynamic cycle, the binding free energy is estimated. Adapted from Ghidini et al. (2025) [21].

Path-based methods. Unlike alchemical approaches, path-based methods, such as Umbrella Sampling (US) [28], MetaDynamics (MetaD) [29] and Well-Tempered MetaDynamics (WT-MetaD) [30], accelerate rare events along physical binding and unbinding pathways. These rely on the definition of a set of collective variables (CVs) (e.g., distances, angles, etc.) that approximate the reaction coordinate and drive exploration of the underlying free energy surface, thereby accelerating the transitions between different metastable states. The result of path-based calculations is often the Potential of Mean Force (PMF), that is the free energy profile along the selected CV. In contrast with standard MD where the regular potential $V(\mathbf{q})$ results in a Boltzmann-weighted sampling, in path-based methods a bias potential is introduced; for example in US the bias potential is written in the form of a harmonic restraint centered at a specific value of the collective variable, ξ_0:

$$\hat{V}(\mathbf{q}) = V(\mathbf{q}) + \frac{1}{2}k(\xi(\mathbf{q}) - \xi_0)^2 \tag{1.4}$$

The bias enhances sampling in the selected regions of phase-space, allowing the system to overcome high free energy barriers. Then, to recover the unbiased Boltzmann

distributions and determine the PMF (see figure 1.2), various post-processing methods have been developed. Among those, WHAM (Weighted Histogram Analysis Method) is the most widely used [31, 32]. This yields both the absolute ΔG and high-resolution mechanistic insights into the transition states.

However, applying path-based methods to cyclic peptides introduces several computational challenges that, to a lesser extent, already affect small molecules. Small molecules typically have low molecular weights (<700 Da) and a limited number of rotatable bonds, meaning their conformational space is relatively confined. Still, capturing the binding and unbinding processes is not trivial: even for small ligands, accurately converging the free energy profile requires a careful study to account for hidden slow degrees of freedom [33]. Cyclic peptides extend these issues considerably. They are typically 7 to 15 residues long and exhibit a large conformational space and, despite backbone cyclization, retain a high degree of residual flexibility, that let them transition between different metastable states separated by high energy barriers [34, 35]. Then the

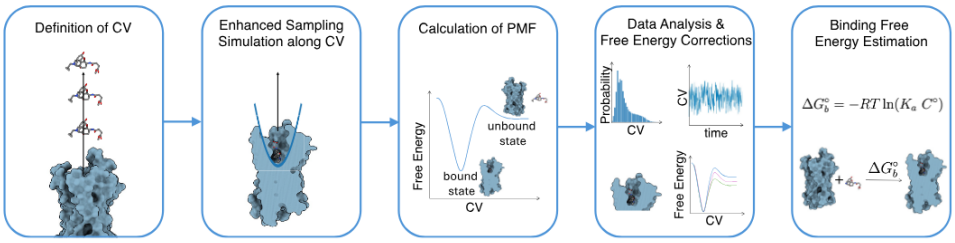


Figure 1.2: Schematic representation of a workflow for calculating absolute binding free energies using path-based methods. After defining a suitable collective variable (CV), enhanced sampling simulations force the exploration of the bound and unbound states, allowing the extraction of the PMF and the estimation of ΔG_b°. Adapted from Ghidini et al. (2025) [21].

main limitations of path-based methods are their sensitivity to the choice of CVs and the well-known "curse of dimensionality" [36, 37]. If the chosen CVs fail to account for all the relevant slow, hidden degrees of freedom, such as internal conformational shifts of a cyclic peptide or the slow rearrangement of water on the surface of the protein, the sampling becomes quasi-nonergodic [38]. This can lead to kinetic trapping and inaccurate free energy estimates. That is the reason why, while rigorous path-methods are really powerful, their application to protein-cyclic peptide complexes has remained unexplored territory in computational biophysics, as we need to develop highly advanced, multi-replica algorithms able to mitigate these dimensional limitations.

1.2.2 The Force Field Dichotomy: Over-compaction vs. Under-binding

Even if enhanced sampling algorithms successfully overcome the kinetic barriers associated with huge conformational transitions, the thermodynamic accuracy of the resulting free energy landscape is bounded by the underlying physical model. In classical molecular dynamics (MD), it is fundamental to determine the time evolution of the system's energy as a function of its atomic coordinates. The gradient of the energy function is related to the forces which are acting on individual atoms. The set of potential energy functions from which the forces are derived is referred to as a *force field*. After decades of improvements, standard additive force fields (such as those belonging to the AMBER, CHARMM, GROMOS and OPLS families) have reached a mature state and are

now widely used to study protein folding and unfolding, protein-protein interactions, enzymatic-reaction mechanisms and ligand binding [39]. However, most force fields efforts in biomolecular simulations have focused on the parametrization of the protein, while less attention was put on the underlying solvent model. The latter is particularly relevant for simulations of intrinsically disordered proteins and peptides (IDPs) for which energy differences between different conformations are small and interactions with water are enhanced [40]. In fact, when simulated with traditional force fields, disordered states are often unstable and tend to collapse into overly compact "molten globule" structures that are nearly as compact as folded proteins, directly contradicting experimental observations [41]. This behaviour is largely attributed to the water models paired with these force fields, which can significantly underestimate London dispersion interactions (in particular water-water and water-solute interactions) [42]. Obtaining the correct balance between these interactions is an important factor to accurately describe the solvation of proteins and may have important consequences for the ability of simulations to correctly describe the highly hydrated disordered states. To improve these deficiencies, researchers developed alternative water models, such as TIP4P-D, that greatly improve the description of disordered protein states, allowing for comparisons between atomistic simulations and experimental studies of such states [42]. Furthermore force fields like a99SB-disp were specifically parameterized to bridge the gap between ordered and disordered states. Through optimization of the backbone torsion potential and of the Lennard-Jones interactions, a99SB-disp was able to achieve high accuracies for simulations of disordered proteins while obtaining accuracies comparable with the best force fields for folded proteins [43]. Despite these remarkable advancements, modeling short, flexible sequences, such as cyclic peptides remains difficult due to their environmental sensitivity: small changes in sequence, solvent or receptor can trigger large conformational shifts. Moreover, unlike globular proteins, which are stabilized by networks of non-bonded interactions, peptides often lack extensive stabilizing contacts. As a result, even small inaccuracies in the force field can lead to significant errors [44–46]. A recent comprehensive benchmark evaluating 11 modern force fields across different peptide systems revealed a functional dichotomy: traditional force fields succeed in maintaining the native structure of rigid miniproteins but fail to capture structural details of disordered proteins. On the other hand, force fields heavily parameterized for IDPs (as ff99IDPs or OPLSIDPSFF) often fail when applied to structured macrocycles [46]. Because no single force field performs best across all systems, researchers are left with the following dichotomy: a model that accurately captures the bound state may artificially over-compact the unbound state, while a model designed to properly solvate the unbound state may lead to an under-binding between the protein and the ligand.

1.3 Aims and Structure of the Thesis

The previous sections (1.2.1, 1.2.2) have presented two challenges that currently limit the computational prediction of binding free energies for protein-cyclic peptide complexes. On one hand, we have seen that rigorous path-based methods may estimate binding free energies with good accuracy, however their performance is heavily dependent on the choice of CVs and is affected by the "curse of dimensionality" [37, 38]. On the other hand, even when sampling is great, the thermodynamic accuracy of the underlying free energy landscape is bounded by the force field we use [43, 46].

The present work addresses these two limitations at the same time. The first aim

is methodological: we develop and validate a multi-replica enhanced sampling protocol capable of getting accurate binding free energies for different protein-cyclic peptide complexes within accessible simulation timescales. To do that, funnel metadynamics [47] is combined with OneOPES [33, 48], within a Hamiltonian Replica Exchange (HREX) architecture [49] implemented through GROMACS [50] and PLUMED [51]. The protocol progressively adds CVs across eight replicas, combining the main collective variables describing the binding process (the funnel projection `pp.proj` and a native contact map, `cmap`) with auxiliary hydration collective variables. Protein hydration is described through the coordination number of water around three virtual atoms, which are built as the center of mass of four protein atoms. Peptide hydration is instead described through collective variables selected using the local fingerprint approach [52], extended to automatically identify "bulk" and "anti-bulk" atoms along the cyclic peptide sequence, following a recent application of hydration collective variables to host-guest systems [53].

The second aim is diagnostic and predictive. Starting from the benchmark dataset introduced by Zhao *et al.* [22], from which the 7K2* protein-peptide family and the structurally distinct, disulfide-constrained 1SFI complex are drawn, we systematically quantify how the intrinsic flexibility of a cyclic peptide correlates with the direction and magnitude of the error introduced by a given force field. We formalize a three-level screening protocol moving from a qualitative, sequence-based assessment, to an explicit-solvent replica-exchange molecular dynamics (REMD) simulation to separate distinct structural families, down to an implicit-solvent REMD simulation within a single family, that assigns the peptide to the force field, or a combination of them, expected to minimize its systematic bias before any production level multi-replica simulation is performed.

Finally, we take a first step beyond this binary force field selection and explore whether the dichotomy between rigid and disordered-state force fields can be addressed directly within the multi-replica scheme itself. To this end, the eighth replica is substituted with one where the cyclic peptide is isolated from the protein in the unbound state through a lower restraining wall, while directly biasing internal conformational CVs to counteract unbound state over-compaction given by the classical force field Amber14SB. The preliminary results obtained with this approach, together with practical limitations encountered, such as the poor overlap between the seventh and eighth replica, are discussed in section 3.5 as an alternative approach to the empirical screening protocol.

The remainder of the thesis is organized as follows. Chapter 2 describes in detail the biological systems, the multi-replica simulation protocol, the fingerprint approach and the thermodynamic analysis pipeline used to assess convergence. Chapter 3 presents the performance of force fields Amber14SB and a99SB-disp on the 7K2* family, the resulting force field dichotomy, the development of the three-level screening protocol, and its resulting application to show how it can improve results, the application of the multi-replica protocol to a structurally different complex (1SFI) and the preliminary results obtained with the extended multi-replica scheme. Finally, conclusions are drawn in Chapter 4.

Chapter 2

Methods

2.1 Computational Techniques and Free Energy Engines

In this section we present the computational framework utilized to study protein-cyclic peptide complexes. We have seen that accurately predicting absolute binding free energies requires overcoming both timescale limitations of classical simulations (since binding and unbinding are rare events) and intrinsic biases of the force fields. Because of that, our methodology integrates several advanced computational tools. We rely on Molecular Dynamics (MD) simulations to capture the conformational behaviour of cyclic peptides in full atomic detail. These simulations are driven by GROMACS [50], which provides an efficient, parallelized engine for molecular simulations. Since forces in MD simulations are calculated using molecular mechanics force fields, we evaluate two state-of-the-art force fields: Amber14SB and a99SB-disp. Then, to actually simulate the binding-unbinding processes, we must resort to enhanced sampling techniques as classical MD timescales are too short to describe these kinds of processes. These methods are powered by PLUMED [51], an open-source plugin that interfaces directly with GROMACS to compute collective variables (CVs) and apply bias potentials on the fly. Through PLUMED, we combine two powerful enhanced sampling methods: OneOPES [48], a replica exchange approach that integrates multi-replica simulations with On-the-fly Probability Enhanced Sampling (OPES) [54–56] to accelerate phase space exploration and Funnel Metadynamics [47], a technique that applies a funnel-shaped potential to restrict the conformational space available to the peptide in the unbound state. In such a way, this approach enhances sampling of both the bound and unbound states. The following subsections will detail each of these components, which build the theoretical and practical foundation for the multi-replica setup used in this thesis and presented in section 2.3.

2.1.1 Molecular Dynamics, GROMACS and Force Fields Overview

Molecular Dynamics (MD) simulations predict the spatial position of each atom as a function of time by calculating the force exerted on each atom by all other atoms and subsequently numerically integrating Newton's equations of motion. The resulting trajectory is a three-dimensional movie that describes the atomic-level configuration of the system at every point during the simulation, allowing us to directly observe conformational fluctuations, solvent dynamics and the binding and unbinding of a peptide to a protein [57]. The forces in these simulations are computed from the gradient of a force field, which is an empirical potential energy fit to the results of quantum mechanical

calculations and to certain experimental measurements. Such force fields are inherently approximate, meaning that their choice is one of the main sources of systematic error in an MD study [57].

All the simulations presented in this thesis were performed in GROMACS [50], an open source MD engine that efficiently distributes all interactions across available CPUs and GPUs; moreover, it is often used for metadynamics together with PLUMED (see A.1), which is really useful for enhanced sampling simulations. GROMACS also supports, among other things, the two solvation regimes that have been used in this work:

- **Explicit solvent**, where every water molecule surrounding the solute is represented atomistically. There are several explicit solvent models (e.g., TIP3P, TIP4P) that provide the most realistic description of hydrogen bonding, ion distribution and solvent reorganisation around the peptide and protein, which is fundamental to accurately estimate binding free energies [58]. The main drawback is that the computational cost is significant, as the number of solvent atoms exceeds the number of solute atoms by more than an order of magnitude.

- **Implicit solvent**, where the solvent is treated as a continuous dielectric medium. This approach reduces computational cost by averaging solvent effects spatially and neglecting altogether solvent-solvent interactions.

The choice between the two models is not only a computational one: as highlighted in section 3.2, they are both important in the development of the screening protocol used in this thesis, since they give us different information on several aspects of a cyclic peptide's conformational behaviour.

Force fields overview. Over the past decades, several force fields families (AMBER, CHARMM, GROMOS, OPLS are the most widely used) have been refined on globular proteins and only recently they have been extended to describe intrinsically disordered proteins and peptides. Here, we introduce the two force fields (Amber14SB and a99SB-disp) that we use for all production runs.

- **Amber14SB.** Amber14SB is a member of the AMBER family, representing an evolution of Amber99SB, using a new fit of all backbone and side chain dihedral parameters to improve deficiencies and biases of previous Amber force fields on secondary structures [59]. It is intended to be used with Transferable Intermolecular Potential with 3 points (TIP3P) but its parametrization was not much concerned, as most of the traditional force fields, with disordered proteins and peptides [60]. As we have seen in section 1.2.2 force fields like Amber14SB can produce an artificially compact unbound ensemble in flexible peptides, translating into a systematic over-binding, since the unbound state is thermodynamically more stable than it should be.

- **a99SB-disp.** a99SB-disp [43] was developed specifically to correct the over-compaction issue introduced by traditional force fields. It is paired with a modified TIP4P-D-like water model (also called a99SB-disp water) and it significantly improves the description of the disordered conformations sampled by flexible peptides in the unbound state but it produces an opposite bias: it tends to destabilise the compact conformations of more rigid macrocycles, leading thus to a systematic under-binding of the more pre-organized complexes.

2.1.2 OneOPES: On-the-fly Probability Enhanced Sampling

Enhanced sampling methods that accelerate a system's dynamics along a set of collective variables (CVs) work as long as we are able to build good CVs. As we have seen in section 1.2.1 if the chosen CVs fail to encode all relevant slow degrees of freedom, convergence will require prohibitively long simulation times. Another class of methods addresses this timescale problem through parallelism, that is simulating multiple replicas of the system at the same time [61–64], while periodically attempting to exchange their coordinates, given a specific physical acceptance criterion. However, since in these methods there is no CV-defined direction, simulations can explore less relevant high-energy portions of phase space, thereby reducing overall efficiency [48]. OneOPES was developed to combine the strengths of both approaches in a single multi-replica framework built around On-the-fly Probability Enhanced Sampling (OPES) [54].

OPES theoretical foundations. In standard OPES, the algorithm aims to sample a target probability distribution $p^{\mathrm{tg}}(\mathbf{s})$, once the method reaches convergence, by considering the following bias potential:

$$V(\mathbf{s}) = \frac{1}{\beta} \log \frac{P(\mathbf{s})}{p^{\mathrm{tg}}(\mathbf{s})} \tag{2.1}$$

where $P(\mathbf{s}) \propto \int d\mathbf{R}\ e^{-\beta U(\mathbf{R})} \delta(\mathbf{s} - \mathbf{s}(\mathbf{R}))$ is the unbiased probability distribution, $\beta = \frac{1}{k_B T}$, $\mathbf{R}$ denotes the atomic coordinates, $U(\mathbf{R})$ the interatomic potential and $\mathbf{s} = \mathbf{s}(\mathbf{R})$ is a set of collective variables. Interestingly the unbiased probability density can be rewritten as an average over the biased ensemble:

$$P(\mathbf{s}) = \frac{\langle \delta(\mathbf{s} - \mathbf{s}(\mathbf{R})) e^{\beta V(\mathbf{s})} \rangle_V}{\langle e^{\beta V(\mathbf{s})} \rangle_V} \tag{2.2}$$

In this way the free energy surface (FES) can be reconstructed using:

$$F(\mathbf{s}) = -\frac{1}{\beta} \log P(\mathbf{s}) \tag{2.3}$$

The typical choice for the target distribution is the well-tempered one, $p^{\mathrm{tg}}(s) \propto [P(\mathbf{s})]^{1/\gamma}$, where $\gamma > 1$ is called the bias factor. Then we only need to compute $P(\mathbf{s})$ through reweighting in order to calculate the bias. OPES builds an on-the-fly estimate of the unbiased probability distribution, by periodically depositing Gaussians, using a weighted Kernel Density Estimation (KDE):

$$P_n(\mathbf{s}) = \frac{\sum_k^n w_k G(\mathbf{s}, \mathbf{s}_k)}{\sum_k^n w_k} \tag{2.4}$$

where the weights depend on the previously deposited bias potential, $w_k = e^{\beta V_{k-1}(\mathbf{s}_k)}$, while $G(\mathbf{s}, \mathbf{s}') = h \exp\big[-1/2(\mathbf{s} - \mathbf{s}')^T \Sigma^{-1} (\mathbf{s} - \mathbf{s}')\big]$ are the Gaussian Kernels, with diagonal covariance matrix $\Sigma_{ij} = \sigma_i^2 \delta_{ij}$ and fixed height $h = \prod_i (\sigma_i \sqrt{2\pi})^{-1}$. The bias potential at step n is thus written as:

$$V_n(\mathbf{s}) = \left(1 - \frac{1}{\gamma}\right) \frac{1}{\beta} \log \left(\frac{P_n(\mathbf{s})}{Z_n} + \epsilon \right) \tag{2.5}$$

where ϵ is a regularization term, that ensures the argument of the logarithm is positive but also allows one to set a limit on the bias, providing better control over desired

exploration. It can be set to $\epsilon = e^{-\beta\Delta E(1-1/\gamma)}$, where ΔE is the height of the free energy barriers one wishes to overcome. Finally, another important component is Z_n, which is a normalization factor computed only over the volume of the CV space explored up to step n:

$$Z_n = \frac{1}{|\Omega_n|}\int_{\Omega_n} P_n(\mathbf{s})\, d\mathbf{s} \tag{2.6}$$

Exploratory Variants. When suboptimal CVs are used, OPES, as every other adaptive-bias enhanced sampling method, is subject to an exploration-convergence tradeoff: it focuses more on quickly converging to a quasi-static bias to provide an accurate free energy estimate and less on exploring the phase space. To prioritize rapid phase space exploration over fast convergence, the OPES Explore variant was introduced [56]. In this variant, instead of writing the bias as a function of the unknown equilibrium distribution ($P_n(\mathbf{s})$), it is written as a function of the on-the-fly estimate that is being sampled in the biased simulation:

$$p_n^{\mathrm{WT}}(\mathbf{s}) = \frac{1}{n}\sum_k^n G(\mathbf{s}, \mathbf{s}_k) \tag{2.7}$$

Then the bias is written as follows:

$$V_n(\mathbf{s}) = (\gamma - 1)\frac{1}{\beta}\log\left(\frac{p_n^{\mathrm{WT}}(\mathbf{s})}{Z_n} + \epsilon\right) \tag{2.8}$$

The modified bias drives the system more aggressively out of metastable states, giving a faster exploration. In addition to OPES Explore, the OneOPES framework includes another variant, called OPES MultiThermal. This variant enhances the fluctuations of the potential energy, allowing the system to sample the multicanonical ensemble corresponding to the range of temperatures $[T_{\min}, T_{\max}]$. OPES MultiThermal acts in a similar way to parallel tempering [63], effectively heating and cooling the system and lowering kinetic barriers along unbiased degrees of freedom.

Multi-Replica Architecture. OneOPES uses different variants across a small set of interacting replicas (we use eight in this thesis, as in the original implementation) as it can be seen in the figure 2.1 below.

Replica 0 contains only OPES Explore bias-potential and it is the most convergence-focused replica. Moreover, since it does not contain temperature scaling, it is used to extract equilibrium properties after reweighting. With increasing replica indices, two more OPES layers are added: OPES MultiCV, which is a set of additional OPES Explore biases applied to transversal CVs and OPES MultiThermal, which is included in higher replicas. Conformations are exchanged across adjacent replicas through coordinate exchanges, which are periodically attempted and are regulated through the Metropolis-Hasting criterion:

$$\alpha = \min\left(1, e^{\frac{-\bar{U}_i(x_j)+\bar{U}_i(x_i)}{k_B T_i} + \frac{-\bar{U}_j(x_i)+\bar{U}_j(x_j)}{k_B T_j}}\right) \tag{2.9}$$

where $\bar{U}(x_i) = U(x_i) + V_i^{\mathrm{TOT}}(x_i)$ represents the total potential energy of replica i and $V_i^{\mathrm{TOT}}(x_i)$ is the sum of all bias potentials applied to it. Then, looking at figure 2.1, this setup creates a double gradient along the replicas: an exploration gradient from replica 0 to 7, that generates new transitions and a convergence gradient from replica

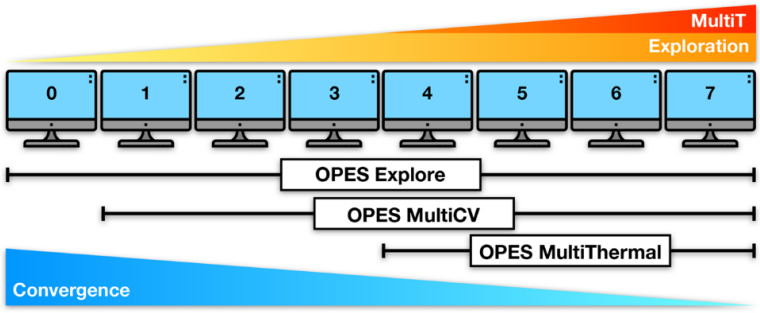


Figure 2.1: Schematic representation of the OneOPES replica exchange method. Replica 0 includes only OPES Explore bias-potential and it is the most convergence-focused replica. Replica 7 includes extra OPES Explore potentials and OPES MultiThermal and it is the most exploration-focused replica. Adapted from Rizzi *et al.* (2023) [48].

7 to 0, that transfers the new conformations back to replica 0. Finally, note that the setup presented in figure 2.1 is not strict and in fact we will slightly modify the order of OPES variants in section 2.3.

2.1.3 Funnel Metadynamics for Absolute Binding Free Energy

Path-based methods like Metadynamics [29] and its evolutions (like OneOPES [48]) accelerate rare events by depositing a history-dependent bias along selected CVs and then reconstruct the Free Energy Surface (FES) through reweighting. When we apply the procedure to binding-unbinding processes (such as protein-cyclic peptide binding), these approaches suffer from a rather simple limitation: once the ligand has left the binding site, it wanders in different solvated states (since they represent a large portion of configuration space), as it can't find its way back to the bound state [47]. This is an important problem, since to accurately estimate binding free energies multiple binding-unbinding events have to be observed.

Funnel Metadynamics. Funnel metadynamics (FM) overcomes this limitation by applying a funnel-shaped restraining potential to the entire system, thus reducing the space to explore in the unbound state and enhancing sampling both in the bound and in the unbound state. Specifically, the potential combines a cone restraint, which includes the binding site, and a cylindrical part, that covers the solvent. As can be seen in figure 2.2 there are four parameters that can be tuned to define the shape of the funnel: the axis z, that defines the exit along the binding path, z_{cc}, that is the distance where the restraint potential goes from the cone shape to a cylindrical one, α, which defines the amplitude of the cone, and R_{cyl}, that is the radius of the cylindrical section.

During the simulation, if relevant parameters are set correctly, the system does not feel any restraining bias while inside the funnel, meaning that the sampling of the bound state is not affected by its presence; only when the ligand reaches the edge of the funnel, a repulsive bias is applied to avoid that the ligand visits regions outside of it. Moreover, because the funnel restraint does not apply any bias during the path from bound to unbound state, the free energy difference between the bound and unbound state, ΔG, exclusively depends on the free energy difference between these two end states. Obviously, to obtain the absolute protein-ligand binding free energy, one needs

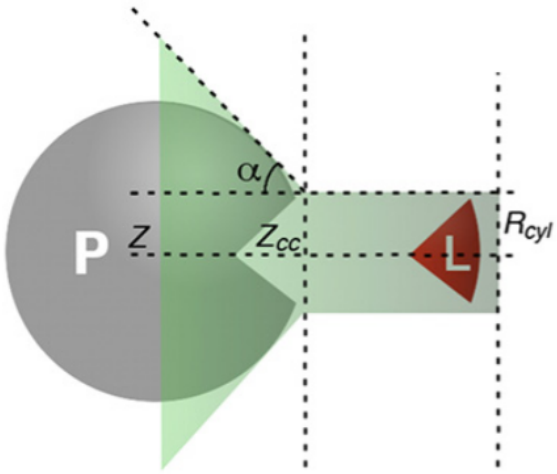


Figure 2.2: Schematic representation of the ligand/protein binding process and the funnel restraining potential. Adapted from Limongelli *et al.* (2013) [47]

to apply an analytical correction that accounts for the restricted cylindrical volume explored in the unbound state relative to the standard-state reference volume:

$$\Delta G_b^\circ = \Delta G - \frac{1}{\beta} \ln \left(\pi R_{\text{cyl}}^2 C^\circ \right) \tag{2.10}$$

where ΔG is the free energy difference between bound and unbound state (see eq. 2.15), computed by the path-based method (e.g., OneOPES), while the second term is the analytic correction and πR_{cyl}^2 represents the surface of the cylinder used as restraint potential. Finally, it is important to note that funnel metadynamics can be combined with other enhanced sampling approaches (as stated by Limongelli *et al.*), which is precisely what we do in this thesis by combining funnel metadynamics and OneOPES.

2.2 Biological Systems

To address the challenges outlined in section 1.3 we require a benchmark of protein-cyclic peptide complexes for which experimental binding affinities are available, so that we can directly compare our computational prediction with the experimental results. In this section we introduce the two biological systems selected for this purpose.

The first set of systems is the 7K2* family [65]: a set of cyclic heptapeptides that are constructed from the DxETGE motif (which are the key interacting residues) of the transcription factor Nrf2 bound to the Kelch domain of human KEAP1. Within this family, the protein target and the overall macrocyclic topology is held fixed, while the peptide sequence is modified by a small number ($\sim 1-2$) of point substitutions. This design can be useful to isolate how small differences in peptide sequence translate into differences in conformational flexibility and how these differences translate into force field dependent errors in the estimated binding free energy. The second system is 1SFI [66] and it is studied as a first validation of the multi-replica and screening protocols developed. It involves a completely different protein target (bovine β-trypsin) and a structurally distinct cyclic peptide: a larger, more rigid macrocycle, further stabilised by a single internal disulphide bond.

Both systems, together with their associated experimental binding affinities, are drawn from the benchmark dataset put together by Zhao *et al.* [22], who compiled a set of protein-cyclic peptide complexes with the respective binding affinities and high-resolution structures in order to benchmark MM/PBSA-based scoring methods. Here we use the same complexes to test our multi-replica protocol, converting the reported binding affinities into binding free energies through equation 1.1.

2.2.1 The Target Dataset: 7K2* Protein-Peptide Complexes

The 7K2* family was built from the work of Ortet *et al.* [65] to recapitulate, in a minimal cyclic peptide, the full binding affinity of Nrf2 for KEAP1. As already introduced in the previous section 2.2, the KEAP1/Nrf2 interaction is mediated by a short recognition sequence on Nrf2, the binding motif DxETGE (Asp77-x-Glu79-Thr80-Gly81-Glu82), which binds KEAP1 in a β-hairpin conformation with a dissociation constant of $K_D \sim 10$ nM. This relatively high binding affinity can be recapitulated in Nrf2-derived linear peptides of ≥ 9 residues that encompass the sequence L**DxETGE**FL [67].

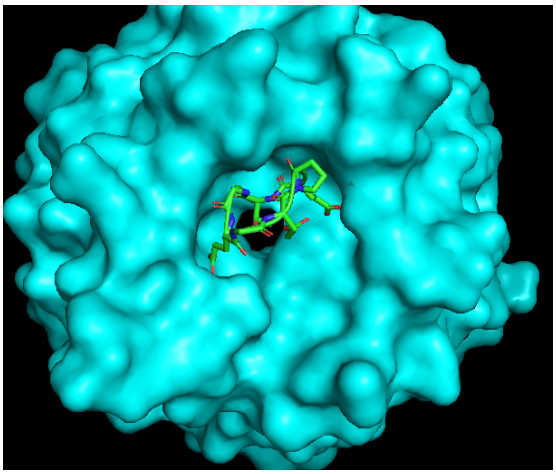


Figure 2.3: Structure of KEAP1 Kelch domain (in cyan) in complex with cyclic peptide 7K2M (in green). The view highlights the characteristic central tunnel of the β-propeller domain, with the peptide bound at the protein surface.

In order to get a cyclic mimetic of this motif, Ortet *et al.* replaced Leu76 with a glycine linker, joining the N- and C-termini of the peptide through a head-to-tail bond while keeping the six residues of the core Nrf2 binding motif, thus yielding the seven-residues macrocycle c[GDEETGE]. This wild-type cyclic peptide (PDB entry 7K2E) bound KEAP1 with $K_D = 470$ nM, which is a 9-fold improvement over the respective linear WT 7-mer but still far weaker than the 33 nM affinity of the 9-residues linear peptide from which it derives. A combination of NMR, X-ray and machine learning analyses of the bound and unbound conformations shows that the gap in binding affinity arises from conformational strain: what emerges is that the cyclization of the 7-mer peptide results in a conformation that is not optimal for binding to KEAP1, meaning that binding requires the macrocyle to be deformed into a high-energy state. Based on this observation, they optimized the conformation of the peptide in the unbound state by introducing point substitutions at three positions within the ring (Asp77, Glu78, Thr80) together with a series of alternative linkers (in place of Gly76) in order to find a balance between conformational pre-organization and residual strain. The optimization

effort produced over 20 cyclic and linear variants, including the cyclic peptide c[DhA-DPETGE] (PDB entry 7K2S), with $K_D = 20 \pm 2$ nM matching the affinity of the linear 9-mer peptide.

In the present thesis, we select six members of this family (composed of only natural residues), presented in table 2.1. These peptides share an identical macrocyclic

System	$\Delta G_{\mathbf{exp}}$ (kcal/mol)	Peptide sequence
7K2E	-8.69	GDEETGE
7K2F	-5.49	GAEETGE
7K2G	-4.85	GDEEAGE
7K2H	-9.41	GDPETGE
7K2I	-6.71	GAPETGE
7K2M	-9.37	GEPETGE

Table 2.1: Systems under study with their respective experimental binding free energies and cyclic peptide sequences.

topology, same 7-residues ring size, same Gly76 linker, and differ only through single or double point substitutions. These characteristics make them a well-controlled benchmark, as any difference in the accuracy of a given force field across these six systems can be attributed to the point substitutions. Specifically, three out of six variants (7K2H, 7K2I and 7K2M) include the Glu78→Pro78 substitution (E78P), 7K2F and 7K2I lack the Asp77 side chain (D77A) and 7K2G lacks the Thr80 side chain (T80A). The substitution pattern is really important, as we can anticipate how the intrinsic flexibility of the unbound peptide, and consequently the expected performance of a given force field, should vary across the different complexes. There are a few points particularly relevant:

- Ortet *et al.* explicitly attributed part of the affinity gain associated with the E78P substitution to increased conformational pre-organization of the free peptide toward the KEAP1-bound geometry. Then, the three peptides (7K2H, 7K2I, 7K2M) containing this substitution are expected to sample a narrower (single-basin) conformational ensemble in the unbound state with respect to the peptides that do not contain the E78P substitution, making them relatively rigid cyclic peptides.

- On the other hand, the three peptides (7K2E, 7K2G, 7K2F) that retain the native glutamate (Glu78) are expected to access a broader (possibly multi-basin) unbound state conformational ensemble. This could be particularly evident in 7K2F and 7K2G, which do not have stabilizing Asp77 or Thr80 side chains that have been shown to play a relevant role in the pre-organization of the free peptide toward the bound state.

These observations will be particularly important when developing the screening protocol to assign to each complex the force field (Amber14SB, a99SB-disp or a combination of them) that minimizes the systematic bias that has been presented in section 1.2.2. Finally, looking at table 2.1, we can see that the six selected variants span a wide range of experimental binding free energies, from 7K2H ($\Delta G = -9.41$ kcal/mol) to 7K2G ($\Delta G = -4.85$ kcal/mol) which is a range wide enough to test the correlation between predicted and experimental binding free energies.

2.2.2 First Validation System: 1SFI

The second system selected for a first validation of the multi-replica and screening protocols is 1SFI, the crystal structure of the sunflower trypsin inhibitor SFTI-1 bound to the bovine β-trypsin [66]. SFTI-1 was isolated from sunflower seeds and characterised by determining its three-dimensional structure in complex with bovine β-trypsin. It is a 14 residues cyclic peptide of sequence c[GRCTKSIPPICFPD], showing a head-to-tail cyclization and further constrained by a single disulphide bond between Cys3-I and Cys11-I, which divides the peptide into a nine-residue loop (the "reactive loop") and a five-residue turn (the "cyclic loop").

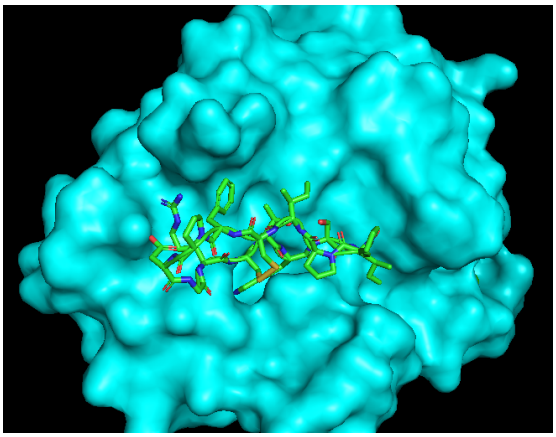


Figure 2.4: Surface representation of bovine β-trypsin (cyan) in complex with the cyclic peptide SFTI-1 (green sticks), highlighting the extended interface between the peptide and the enzyme active-site.

Unlike the 7K2* family, SFTI-1 is not an engineered mimetic of a linear binding motif but a natural peptide belonging to the Bowman-Birk inhibitor (BBI) structural class, from which it differs quite a bit as it is much shorter (BBI are typically around 60-70 residues long) and it is cyclic instead of linear. Despite its size SFTI-1 is an extremely potent inhibitor of bovine β-trypsin, with a dissociation constant of $K_D = 0.1$ nM ($\Delta G = -13.73$ kcal/mol). The high value of K_D has been attributed to the significant conformational rigidity obtained through the cyclic nature of the peptide and the presence of the internal disulphide bond. These characteristics lead to a good pre-organization of the peptide into a bound-like conformation which, in turn, should result in a small entropic penalty upon binding, that is also consistent with the extensive solvent-accessible surface area occluded at the inhibitor-enzyme interface.

Comparison with the 7K2* family. 1SFI has been chosen as a structurally distinct system with respect to the 7K2* family described in the previous section 2.2.1, with the aim to provide a first test of the generality of the multi-replica and screening protocols (developed in sections 2.3 and 3.2). There are several differences between the two systems which are relevant for the developed protocols:

- **Size and topology.** The 7K2* peptides are head-to-tail cyclic heptapeptides, while 1SFI is a 14-residues head-to-tail macrocycle, additionally constrained by an internal disulphide bond. The presence of this bond makes the two systems topologically different.

- **Protein target.** The 7K2* peptides all bind to the same target, the Kelch domain of human KEAP1, at a solvent-exposed site on the surface of the protein, as it can be seen in figure 2.3. 1SFI binds bovine β-trypsin by inserting the reactive loop into the enzyme's active-site.
- **Intrinsic flexibility.** This characteristic is not directly comparable between the two systems. In fact the E78P substitution which leads to the pre-organization of the free peptide in 7K2* has no counterpart in 1SFI. The rigidity in 1SFI in the unbound state arises both from head-to-tail cyclization and from the internal disulphide bond. This is important in the context of the screening protocol, as we need to understand how rigid the peptides are to assign them the correct force field or force field combination.
- **Binding free energy.** The binding free energy of 1SFI (-13.73 kcal/mol) is more negative (that is, stronger binding affinity) than that of any 7K2* variants, which span from -4.85 to -9.41 kcal/mol, consistent with the more extensive interface and higher degree of pre-organization of the free peptide.

2.3 Multi-Replica Simulation Setup

In this section we describe how PLUMED, OneOPES and funnel metadynamics, which we have introduced in previous subsections (A.1, 2.1.2, 2.1.3), are combined into the production protocol used to run every multi-replica simulation in this thesis. The simulation setup follows the OneOPES architecture described in section 2.1.2: eight replicas with two main collective variables (`pp.proj` and `cmap`), biased in every replica by a common OPES Explore potential, and several layers of hydration CVs and OPES MultiThermal, which are progressively added as the replica index increases. The setup is summarised in table 2.2 below.

Replica	Main CVs	Protein Hydration (V^*)	Peptide Hydration (A^*)	OPES MultiThermal
0	`pp.proj`, `cmap`	–	–	–
1	`pp.proj`, `cmap`	V_1	–	up to 305 K
2	`pp.proj`, `cmap`	V_1, V_2	–	up to 310 K
3	`pp.proj`, `cmap`	V_1, V_2, V_3	–	up to 320 K
4	`pp.proj`, `cmap`	V_1, V_2, V_3	A_1	up to 330 K
5	`pp.proj`, `cmap`	V_1, V_2, V_3	A_1, A_2	up to 340 K
6	`pp.proj`, `cmap`	V_1, V_2, V_3	A_1, A_2, A_3	up to 350 K
7	`pp.proj`, `cmap`	V_1, V_2, V_3	A_1, A_2, A_3, A_4	up to 360 K

Table 2.2: Progressive addition of collective variables and OPES MultiThermal across the eight replicas of the setup. Main CVs `pp.proj` and `cmap` are biased throughout, while protein hydration CVs (V^*) and peptide hydration CVs (A^*) are progressively added. For 1SFI, only V_1 is defined.

Each replica contains the entire complex (protein and cyclic peptide). Replica 0 carries only the main OPES Explore bias on `pp.proj` and `cmap`, it is at $T = 300K$ and consistently with section 2.1.2 it is the most convergence-focused replica and the one used for post-processing analysis (see section 2.4). Replica 7 contains instead all nine CVs and OPES MultiThermal up to $T_{\max} = 360K$ and it is the most exploration-focused replica. Then, coordinate exchanges between neighbouring replicas are really important as exploration-focused replicas can exchange their conformations to lower-index replicas in order to enhance sampling. These exchanges are attempted throughout the simulation and they are regulated by Hamiltonian Replica Exchange (HREX) (see

appendix A.2). We will present now, in detail, the two main CVs shared by every replica.

The funnel projection: `pp.proj`. As we have seen in section 2.1.3, funnel metadynamics requires the definition of an axis z along which the peptide is projected (the exit direction), the switching distance z_{cc}, the cone aperture α and the radius of the cylinder R_{cyl} that together define the shape of the funnel restraint. Here, we define the axis through two reference points: p_0, the center of mass of four protein atoms inside the protein, and p_1, the center of mass of six protein atoms on the solvent-exposed protein surface. These atoms were selected manually in PyMOL for each complex in such a way that the axis $p_0 - p_1$ was aligned to the ideal unbinding pathway of the peptide. Then, `pp.proj` is defined as the projection of the peptide center of mass onto the $p_0 - p_1$ axis. This CV is used both as OneOPES main CV and as the axis z of the funnel restraint. Note that, while we tried to build the most general possible setup, funnel parameters have to be adjusted for complexes of drastically different sizes (as is the case between the 7K2* and 1SFI complexes). Moreover, the manual selection of p_0 and p_1, the choice of protein hydration CVs V^* and the funnel parameters remain, as of now, the only steps in the whole CV-selection pipeline which are not fully automated. Their automation is left for future work but methods to automatically identify the binding pocket of the protein already exist [68–70].

The native contact map, `cmap`. The second main CV, `cmap`, quantifies native protein-peptide interface through the sum of six protein-peptide heavy-atom contacts. Contacts are identified computationally as follows: for each peptide residue, the closest protein heavy atom (within a chosen cut-off distance) is found and exactly six contacts are taken, exploiting the ring topology of the cyclic peptides and maximizing the distance between chosen contacts to avoid highly correlated ones. Each contact k enters `cmap` through the following rational switching function:

$$s(r) = \frac{1 - \left(\frac{r-d_0}{r_0}\right)^n}{1 - \left(\frac{r-d_0}{r_0}\right)^m} \tag{2.11}$$

where $r_0 = 0.55$ nm, $d_0 = 0$, $n = 4$ and $m = 8$. Then $\texttt{cmap} = \sum_{k=1}^{6} s(r_k)$, so that it is close to 6 when all native contacts are formed and decreases smoothly as the peptide unbinds.

2.3.1 The Fingerprint Approach for Solvation

Theoretical background. The hydration collective variables used in the multi-replica setup (see table 2.2) are built using the local fingerprint introduced by Pérez-Conesa *et al.* [52]. For a solute atom i, the fingerprint is defined as follows:

$$S_i = -2\pi\rho_{\text{w,loc}} \int_0^\infty \{g_{iw}(r)\ln[g_{iw}(r)] - g_{iw}(r) + 1\}\, r^2\, dr \tag{2.12}$$

where $\rho_{\text{w,loc}}$ is the local number density of water and $g_{iw}(r)$ is the radial distribution function (RDF) between atom i and water oxygen atoms. Note that the integrand is constructed in such a way that it vanishes in the case of a uniform distribution, $g_{iw}(r) = 1$ (i.e., whenever water at distance r from atom i is distributed as it would be in the absence of interactions). Then, the local fingerprint measures how much

the local arrangement of water around atom i changes with respect to the bulk-like ideal reference ($g_{iw}(r) = 1$). A strongly negative value of S_i can be identified with an atom around which water is heavily perturbed relative to the bulk: this can happen either because the atom is buried (equivalently, it excludes a large volume of solvent), or because the water that manages to be around it is forced in an ordered, hydration shell. Both behaviours are not bulk-like and can be seen as an indication of hydrophobicity. On the other hand, a value of S_i close to zero, can be identified with an atom whose surrounding RDF stays close to 1 and water around it is barely perturbed [52].

Building on the local fingerprint idea, Schulze *et al.* [53] proposed an automated strategy to select hydration CVs, by computing the fingerprint for atoms of a host and a guest molecule and classifying atoms as bulk or anti-bulk depending on, respectively, highest and lowest fingerprint values. After selecting bulk and anti-bulk atoms, they constructed the CVs by computing the water coordination number around these atoms. This application to host-guest systems directly motivates its extension to cyclic peptides developed in this thesis.

Extension to cyclic peptides. To construct the cyclic peptide hydration CVs A^*, we perform unbiased simulations in explicit solvent for 100 ns of the free peptides. The radial distribution function between every heavy atom of the peptide and water oxygens is computed from the resulting trajectory using the `rdf` command of GROMACS up to a radius of 2 nm. Subsequently the fingerprint is evaluated for each atom by numerically integrating 2.12, using the trapezoid rule, which produces a ranking of all heavy-atoms of the peptide from the most anti-bulk (lowest S_i) to the most bulk-like (highest S_i). Four atoms are selected from this ranking: the atom with the lowest and highest values of fingerprint are chosen, respectively, as the first anti-bulk and bulk atoms. The second anti-bulk and bulk atoms are chosen as a compromise between their value of fingerprint and the distance between the atom and the first chosen anti-bulk/bulk atom. In particular, we choose them based on the following criteria:

1. Reject atoms directly connected to the first selected atom.
2. Compare atoms with fingerprint difference < 0.2 with respect to the first chosen atom.
3. If multiple atoms have the same fingerprint value, we choose the one with the longest path connecting it to the first chosen atom.
4. Loop through the entire table of fingerprint values until the fingerprint difference is not < 0.2

Finally, the hydration CVs $A_1, \ldots, A_4$ are defined as the coordination number of water oxygens around the four selected atoms, defined as follows:

$$CN_i = \sum_{j}^{r_{ij}<r_{NL}} \frac{\left(\frac{1-\left(\frac{r_{ij}}{r_0}\right)^n}{1-\left(\frac{r_{ij}}{r_0}\right)^m} - \frac{1-\left(\frac{d_{MAX}}{r_0}\right)^n}{1-\left(\frac{d_{MAX}}{r_0}\right)^m} \right)}{\left(1 - \frac{1-\left(\frac{d_{MAX}}{r_0}\right)^n}{1-\left(\frac{d_{MAX}}{r_0}\right)^m} \right)} \tag{2.13}$$

where r_{ij} is the distance between heavy-atom i and water oxygen j, summed over all the atoms within a sphere of radius $r_{NL} = 1.6$ nm. Moreover $r_0 = 0.35$ nm, $d_{MAX} = 1.2$ nm (beyond which the function is zero), $n = 6$ and $m = 10$.

Protein hydration CVs. The fingerprint approach, as described in [52, 53] cannot be applied to the protein, as the majority of the protein heavy atoms are buried inside the folded structure, meaning these atoms would be assigned an artificially negative, anti-bulk-like fingerprint value. Since this effect would dominate, a fingerprint ranking of protein atoms would be useless. Then, the protein hydration CVs V^* are defined in a much simpler way: we take four real C_α atoms chosen depending on the protein topology, and their center of mass defines a virtual atom, around which we take the water coordination number, using the switching function defined in equation 2.13 with the same parameters used for peptide hydration CVs, except $n = 2$ and $m = 6$. This difference between the peptide and protein hydration CVs arises from the nature of the atoms we have chosen. This is explained in ref. [71]: the CVs need to be sufficiently short range to focus on water molecules near atom i, but they also have to be sufficiently long range to make its derivative smooth with respect to more distant water molecules. For peptide hydration CVs the choice of $n = 6$ and $m = 10$ guarantees a good compromise between these two aspects. For protein hydration CVs, since we use virtual atoms, there is no repulsion between atom i and water molecules, meaning we can have $r_{ij} = 0$. The choice of $n = 2$ and $m = 6$ yields a switching function with the same long range behaviour but with a softer core at short distances, which prevents numerical instabilities.

For the 7K2* family, three virtual atoms are used, on the surface of the protein, at the center of the protein tunnel and at the end of the tunnel. In this way water flows through the tunnel and favours re-binding. 1SFI lacks a protein tunnel and a single virtual atom on the surface of the protein has been used.

2.4 Thermodynamic Analysis Protocol

We describe now, from a methodological point of view, how the output of a OneOPES/funnel metadynamics simulation, that is the biased trajectory and the bias potential deposited along replica 0, are post-processed to extract two types of information:

- A quantitative estimate of the binding free energy ΔG_b°, together with its associated uncertainty.

- A set of measures used to evaluate the sampling efficiency and rigorously judge if a given production run has reached convergence. System-specific choices (i.e., equilibration time, length of time blocks, etc.) are different between complexes 7K2* and 1SFI and are reported both in chapter 3 and appendix F; here we describe the general procedure which is the same for every system.

Two layers of analyses are used. The first one (section 2.4.1) reweights the biased trajectory of replica 0 to reconstruct the one-dimensional free energy profile along `pp.proj` and get the absolute binding free energy ΔG_b°. We compare this estimate across three independent simulations and require their standard deviation to be below a fixed threshold, thus providing a primary convergence criterion. The second one (section 2.4.2) evaluates the stationarity of the entire free energy surface (FES) through the computation of FES-RMSD over cumulative time-blocks, while tracking physical transitions to ensure ergodicity. Neither analysis is, on its own, a sufficient proof of convergence; they are used together and their combination is discussed to assess convergence for every system (see chapter 3).

2.4.1 Multi-Run Convergence: ΔG_b° Integration and Cross-Replica Consistency

Reweighting procedure. For each system, we perform three independent simulations and use the trajectory of replica 0, which is the convergence-focused, unbiased in temperature replica, for thermodynamic analysis. Each frame of the trajectory is assigned to a statistical weight:

$$w = \exp\left(\beta \left(V(\mathbf{s}(\mathbf{X}), t) - V_{\max}\right)\right) \tag{2.14}$$

where $V(\mathbf{s}(\mathbf{X}), t)$ is the total instantaneous bias acting on the frame, while $V_{\max}$ is the highest bias value, which is subtracted for numerical stability. This is the implementation of the weight calculation introduced in 2.1.2, in eq. $w_k = e^{\beta V_{k-1}(\mathbf{s}_k)}$ and 2.2. An initial portion of the trajectory, which is treated as equilibration, is discarded before reweighting. The reweighted trajectory is then projected along `pp.proj` to reconstruct the one-dimensional free energy profile $F(z)$, where $z \equiv$ `pp.proj` is the funnel axis. The binding free energy between the bound region B and the reference unbound state is obtained as follows:

$$\Delta G = -k_B T \ln \left[\int_B dz \ \exp\left(-\beta (F(z) - F_U)\right) \right] \tag{2.15}$$

where F_U is the constant value of $F(z)$ in the unbound region of the funnel. The absolute binding free energy is subsequently obtained by adding the analytical correction term that accounts for the restriction in the unbound state due to the cylindrical funnel (eq. 2.10). The procedure is repeated on progressively longer portions of the trajectory, so to obtain the absolute binding free energy as a function of the simulation time, $\Delta G_b^\circ(t)$. Note that the reweighting procedure is run independently for all three simulations of a given system.

First convergence check. Once we have the estimates $\Delta G_b^\circ(t)$ for all three simulations, we combine them at each point in time by computing the mean and standard deviation; the same combination is applied to both the one-dimensional free energy profile $F(z)$ and consequently to the scalar ΔG_b° so that we can study at the same time the free energy landscape along the funnel projection and the absolute binding free energy as a function of time. The standard deviation of $\Delta G_b^\circ(t)$ across three independent simulations represents the primary quantitative convergence criterion used in this thesis. In particular, we follow the work of Karrenbrock *et al.* [33], where they considered coarse-grained, statistically independent snapshots of the trajectory, sampled every 100 ns after equilibration, with convergence achieved when at least three consecutive points (which are obtained as the mean and standard deviation of the three independent simulations) have a standard deviation of at most 1 kcal/mol. Here, we adapt this strategy to the shorter trajectories available in this work. After discarding the equilibration time (which is the same used for $\Delta G_b^\circ(t)$), the remainder of the trajectory is divided into a small number of windows (of 40/50 ns each, depending on the system), and 2.15 together with 2.10 are evaluated once per each window, yielding a handful of estimates of ΔG_b°, sampled more sparsely than the continuous curve. Then, at each of these points we compute the mean and standard deviation across the three independent simulations and plot them as error bars. Results of these analysis are reported, for each system, in chapter 3. Because of the length of our simulations, this coarser analysis is not meant to establish, on its own, convergence, however we

can understand from it whether the cross-replica standard deviation is decreasing with time and whether the mean estimate is moving toward the experimental value. As the length of simulations is extended in future work, this analysis will be extended to the point of supporting a convergence assessment based on several long, *weakly-correlated* windows.

2.4.2 Single-Run Convergence: FES-RMSD and Sampling Efficiency

The analysis of section 2.4.1 gives back an estimate of the absolute binding free energy ΔG_b° and its cross-replica standard deviation, however this could mask underlying instabilities, especially if the system is trapped in a local minimum. To rule out false-positive convergence and verify the stationarity of the FES, we introduce some metrics based on the RMSD of the free energy profile and on sampling efficiency, which are extracted from the post-equilibration portion of the replica 0 trajectory:

- **Number of binding/unbinding round trips.** The number of bound $\rightarrow$ unbound and unbound $\rightarrow$ bound transitions observed during the whole post-equilibration trajectory. Note that this count is inflated because of replica-exchange; in fact a configuration can switch states in a short amount of time when it is exchanged with a replica whose configuration is in a different state. Then, we cannot really consider this as a measure of the physical binding/unbinding kinetics of a single continuous trajectory but as an indicator of mixing efficiency between replicas.

- **Real time % sampling.** The fraction of simulation time, before reweighting, spent in the bound, unbound and transition states. Comparing these percentages allows us to understand whether the multi-replica setup is effectively driving the system out of the binding pocket and into the unbound state.

- **FES-RMSD analysis.** To prove the stationarity of the free energy profile, we divide the trajectory of replica 0 into cumulative time windows, that is blocks of 40 or 50 ns, depending on the system. For each block, the FES is reconstructed and aligned to zero at its minimum. We compute the RMSD in kcal/mol between the current block and the previous one as follows:

$$\text{FES-RMSD} = \sqrt{\frac{1}{N_B} \sum_{k \in B} \left(F_i(z_k) - F_{i-1}(z_k)\right)^2} \tag{2.16}$$

  where $F_i(z_k)$ and $F_{i-1}(z_k)$ represent the free energy values at the k-th bin of the funnel projection axis z for the current block i and the previous block $i-1$, respectively. The summation is restricted to thermodynamically relevant (one of the two free energy profile $<$ free energy cut-off, which depends on the system) bins B to avoid numerical noise and N_B is the total number of bins in subset B.

These checks are used at the same time to establish a robust convergence criterion. In particular, we consider that the FES has reached a stationary, converged regime when the FES-RMSD between successive blocks remains below the 0.6 kcal/mol target for several windows; the precise number will depend on the length of these windows but regardless our simulations currently are too short to state definitive convergence using any convergence criterion. The target value of 0.6 kcal/mol has been chosen because, at $T = 300K$, the thermal noise of the system is equal to $k_B T \approx 0.6$ kcal/mol, so

any fluctuation of the free energy profile below this threshold cannot be distinguished from the intrinsic thermal noise of the solvent. The tolerance set here is stricter than the 1 kcal/mol threshold set in section 2.4.1 for the standard deviation of $\Delta G_b^\circ(t)$ as in that case noise could average out along the integration to obtain ΔG_b°, while the FES-RMSD is a point-wise comparison across the whole profile so there's no need to increase the safety margin since there is no integration to be performed.

Chapter 3

Results and Discussion

In this chapter we present the results obtained using the methods introduced in chapter 2. We start by presenting the baseline performance of both Amber14SB and a99SB-disp force fields (section 3.1), applied to the 7K2* family, showing that the two isolated force fields fail in opposite directions. Following the resulting force-field dichotomy we introduce a predictive three-Level screening protocol (section 3.2), which assigns, prior to any production run, the force field (or force field combination) expected to minimize the systematic bias obtained with both isolated force fields. Then, in section 3.3, we show the results obtained by applying the screening protocol, which substantially improves the agreement between simulated and experimental binding free energies, reducing the mean absolute error (MAE) to the commonly used 1 kcal/mol target. In section 3.4 we validate our multi-replica and screening protocols to a structurally distinct, more rigid complex outside of the 7K2* family, 1SFI. Results show that the protocol generalizes pretty well for this larger and more rigid system. Finally, in section 3.5, we present preliminary results of an extended multi-replica scheme that attempts to overcome the force field dichotomy.

3.1 The Force Field Dichotomy: Baseline Performance

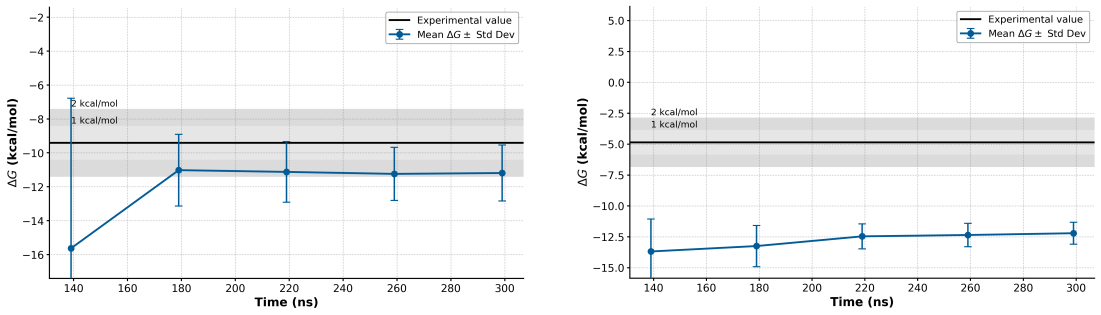
We present now the results obtained by simulating the 7K2* family, all six complexes, independently with Amber14SB and a99SB-disp, using the multi-replica setup presented in section 2.3. For each system we performed three independent simulations of 300 ns, discarding the first 100 ns as equilibration. Computational details about how the simulation is performed and OneOPES parameters are set can be found in appendix F. The purpose of this baseline characterization is to demonstrate that neither Amber14SB nor a99SB-disp are, on their own, able to reproduce the experimental binding free energy for all systems, and that the force fields fail in opposite directions, as anticipated in section 1.2.2. In particular, we will show that these errors are not symmetric: the error introduced by Amber14SB in one direction is considerably larger than the one introduced by a99SB-disp in the opposite direction. This is an interesting result, in light of the hybrid approach presented in section 3.2, since the analysis performed on the 7K2* complexes reveals that a variance-weighted average, instead of the simple arithmetic mean that we use, would have been a more accurate consensus estimator (at least for some systems). Obviously, this observation could have been made only after having the data from the simulations presented in this section and the arithmetic mean has been chosen precisely to avoid this post-simulation tuning.

For each force field, we present a detailed analysis for two representative systems,

7K2H and 7K2G. We choose these two systems because, based on their biological structure (see 2.2.1) and on the Δ_{rmsd} obtained from the implicit solvent REMD simulations, 7K2H is the most rigid and pre-organized peptide of the family, while 7K2G is expected to be the most flexible one. The full set of analyses for the four remaining systems (7K2E, 7K2F, 7K2I, 7K2M) are reported in Appendix G.1 for results with Amber14SB and Appendix G.2 for results with a99SB-disp.

3.1.1 Amber14SB: Over-compaction and Over-binding

Convergence and accuracy. Figure 3.1 shows the absolute binding free energy $\Delta G_b^\circ(t)$ for 7K2H and 7K2G, averaged over three independent simulations and evaluated at five time points, every 40 ns after equilibration, as described in section 2.4.1. For 7K2H, the cross-replica standard deviation never drops below 1 kcal/mol but steadily decreases from 2 kcal/mol at 180 ns to 1.5 kcal/mol at 300 ns. This does not satisfy the primary convergence criterion introduced in section 2.4.1. The mean ΔG_b° is stable around -11.2 kcal/mol, rather close to the experimental binding free energy. For 7K2G, the cross-replica standard deviation decreases from 2.6 kcal/mol at 140 ns to 0.8 kcal/mol at 300 ns. In this case, the standard deviation drops below 1 kcal/mol in the last two windows but we need a more extended simulation to effectively assess convergence. Differently from 7K2H, for 7K2G the mean ΔG_b° is around -12.5 kcal/mol and pretty far from the experimental binding free energy. This result should be expected, following the over-compaction issue presented in section 1.2.2.

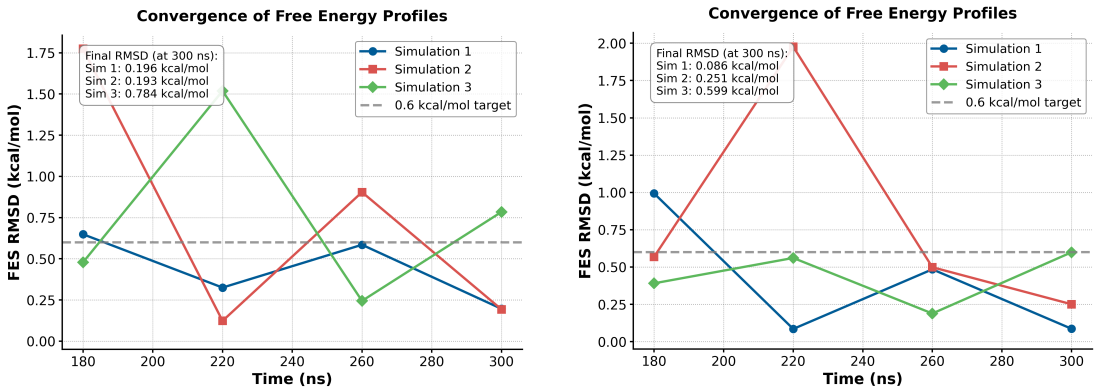


(a) 7K2H (b) 7K2G

Figure 3.1: Absolute binding free energy $\Delta G_b^\circ(t)$ (mean $\pm$ std over three independent simulations) evaluated at five sparse time points following section 2.4.1. Results are shown for 7K2H (a) and 7K2G (b) simulated with Amber14SB. The black lines mark the experimental values.

The behaviour observed in figure 3.1 should be confirmed by the second convergence criterion introduced in section 2.4.2, the FES-RMSD. Figure 3.2 below shows the FES-RMSD between successive 40 ns blocks for three independent simulations. For 7K2H the FES-RMSD is always above 0.6 kcal/mol for at least one of the three independent simulations, however we can observe the downward trend for all three simulations going from 180 ns to 300 ns, confirming the drop in standard deviation already observed in figure 3.1a. For 7K2G, the FES-RMSD reaches a value below the threshold in the last two windows, which confirms the behaviour of the cross-replica standard deviation seen in figure 3.1b.

Figures 3.1 and 3.2 show that 7K2H, which is closer to the experimental ΔG_b°, is less close to convergence with respect to 7K2G. This is an important reminder that



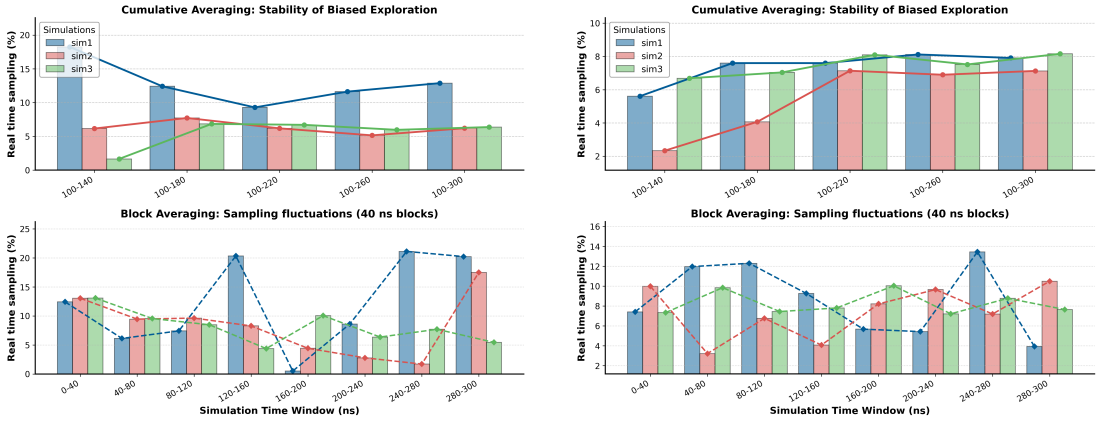
(a) 7K2H (b) 7K2G

Figure 3.2: FES-RMSD between successive 40 ns blocks for the three independent simulations of 7K2H and 7K2G with Amber14SB. The dashed line represent the 0.6 kcal/mol target.

convergence and accuracy are two distinct properties of a simulation: while 7K2H's binding free energy is only ~ 1.8 kcal/mol distant from its experimental value, 7K2G's binding free energy is ~ 7.4 kcal/mol away from its experimental value, showing a heavy over-binding despite excellent sampling statistics, as we can see in the following paragraph.

Sampling efficiency. Figure 3.3 shows the real-time sampling of the unbound state described in section 2.4.2, both as a cumulative average and as a block average over 40 ns windows. Recall that here the bound/transition/unbound states distinction is based only on the value of `pp.proj`, meaning that the bound state is not the native, bound pose but simply a geometric definition. To be precise: if `pp.proj` < 1.7 nm the peptide is in the bound state, if `pp.proj` > 2.5 nm the peptide is in the unbound state, if it is in between these two states, the peptide is in the transition state. These definitions are valid for each one of the 7K2* complexes. We focus on the sampling of the unbound state, since it is the most sensitive indicator of the over-compaction artifact introduced by Amber14SB. We expect that, if Amber14SB over-compacts the peptide in the unbound state, the system struggles to explore it; the comparison of the unbound state exploration between the simulation with Amber14SB and a99SB-disp will be pretty revealing. For 7K2H the unbound state sampling is between 6% and 13% and from the block averaging we can see that it is explored in each window, meaning there are no sampling issues. For 7K2G, we observe almost the same behaviour, with all three replicas aligned around 8%. If we compare this result with what we have obtained for a99SB-disp (in figure 3.9b), there is a clear over-compaction arising for this system due to Amber14SB.

Moreover, global real-time populations (bound, transition, unbound) and number of binding-unbinding events (round trips) are reported in table 3.1 below. As we can see the number of round trips is high in each simulation but, as we have said in section 2.4.2, this number is inflated by replica-exchange. Nevertheless, this shows that there's no bottleneck in the exchange between different replicas. In addition, the 1d Free Energy Surfaces (FES) along the funnel projection (z) are reported in figure 3.4. Both profiles correctly exhibit a deep energy minimum associated to the bound state (note that the native pose is at `pp.proj` $\equiv z \approx 1.47$ nm) and a flat unbound region



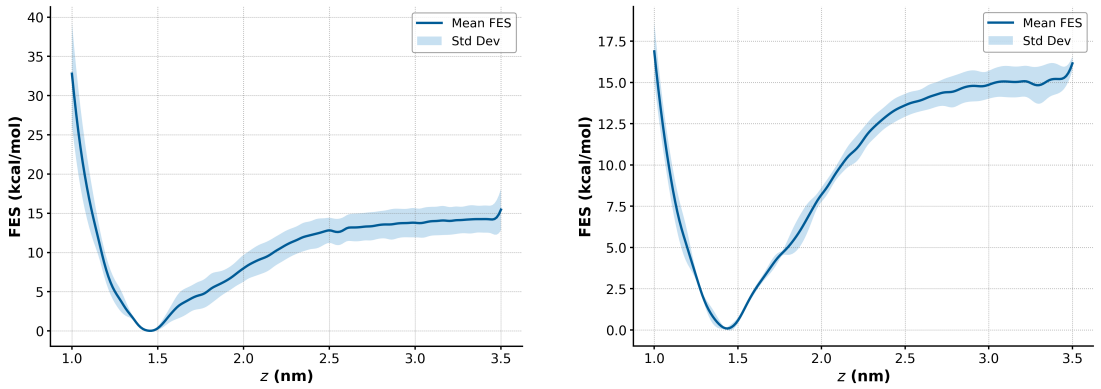
(a) 7K2H (b) 7K2G

Figure 3.3: Cumulative (top) and block-averaged (bottom) real-time sampling of the unbound state of three independent simulations of 7K2H and 7K2G with Amber14SB.

System	Sim	Bound (%)	Transition (%)	Unbound (%)	Round trips
7K2H	1	51.7	35.4	12.9	1160
	2	64.2	29.6	6.2	837
	3	67.0	26.6	6.4	874
7K2G	1	57.5	34.6	7.9	1000
	2	60.8	32.1	7.1	866
	3	57.7	34.1	8.2	973

Table 3.1: Real-time sampling percentages for the bound, transition and unbound states, together with the total number of round trips for systems 7K2H and 7K2G simulated with Amber14SB.

(for $z > 2.5$ nm), which is the expected behaviour for correct bulk sampling in funnel metadynamics. However, particularly for 7K2G, we can see that the free energy associated to the unbound state is high (around 15 kcal/mol) which is an evident sign of the over-compaction in the unbound state produced by Amber14SB. Another measure



(a) 7K2H (b) 7K2G

Figure 3.4: 1d Free Energy Surfaces (FES) along the funnel projection axis (z). The profiles exhibit a characteristic deep energy minimum, associated with the bound state and a flat energy region, associated with the unbound state.

of efficiency of replica-exchange is the exchange probability which ranges between 30% and 65% for 7K2H and between 20% and 60% for 7K2G. In both cases, low-index replicas show a higher replica-exchange probability, while high-index replicas show a lower replica-exchange probability. This is expected, since as the replica index increases, we add more CVs and increase temperature, meaning that high-index replica differ more with respect to low-index replicas.

Correlation with peptide flexibility. Figure 3.5 shows the relation, across all six 7K2* complexes, between the relative error of the Amber14SB estimate ΔG_b° with respect to experiment and the intrinsic flexibility of the free peptide, measured by Δ_{rmsd} (obtained from implicit solvent REMD simulations, see section 3.2). A clear positive trend emerges: Pearson $r = 0.89$, Spearman $\rho = 0.89$ and Kendall $\tau = 0.73$. Because of the limited sample size ($N = 6$), we tested the statistical robustness of this trend against sample variation by performing a bootstrap resampling analysis of 10k iterations. We can see that the 95% confidence intervals (CI) remain strictly positive, demonstrating that this result is not an artifact given by a single outlier. This confirms that the over-compaction of force field Amber14SB paired with water model TIP3P is not uniform across peptide of the same family but it grows with the peptide's flexibility.

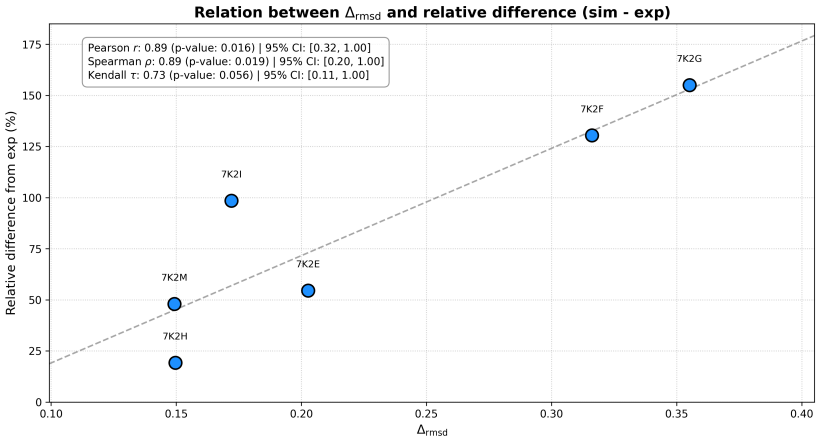


Figure 3.5: Relative error of the ΔG_b° estimate with respect to experiments, as a function of Δ_{rmsd}. Correlation coefficients and 95% bootstrap confidence intervals are reported in the text box.

Overall accuracy. Figure 3.6 shows the correlation between the ΔG_b° estimate and the experimental values across the whole family. In order to rigorously assess the statistical significance of the predictions, the correlation coefficients, mean absolute error (MAE) and mean signed error (MSE) were not computed as single values. We employed a Monte Carlo error propagation approach: we generated 10k datasets where, for each complex, the ΔG_b° was randomly sampled from a normal distribution centered on the mean value of the binding free energy and having σ equal to its cross-replica error. The metrics were evaluated for each of the 10k iterations, allowing us to extract a mean, standard deviation and a 95% confidence interval (CI). We can see from figure 3.6 that there's basically no correlation: $r = 0.09 \pm 0.37$, $\rho = 0.04 \pm 0.34$ and $\tau = 0.08 \pm 0.28$. Moreover, we can see that every system is below the diagonal, reflecting once again the over-binding observed for 7K2H and 7K2G. We can quantify this bias through the mean absolute error and the mean signed error, shown in the text box of figure 3.6. In

particular the MSE $= -5.38 \pm 0.40$ kcal/mol, tells us that the bias is in the over-binding direction and that, on average, its magnitude is over 5 kcal/mol. These results make it clear that the Amber14SB force field cannot, on its own, accurately estimate the binding free energy for all complexes of the 7K2* family with the current multi-replica setup.

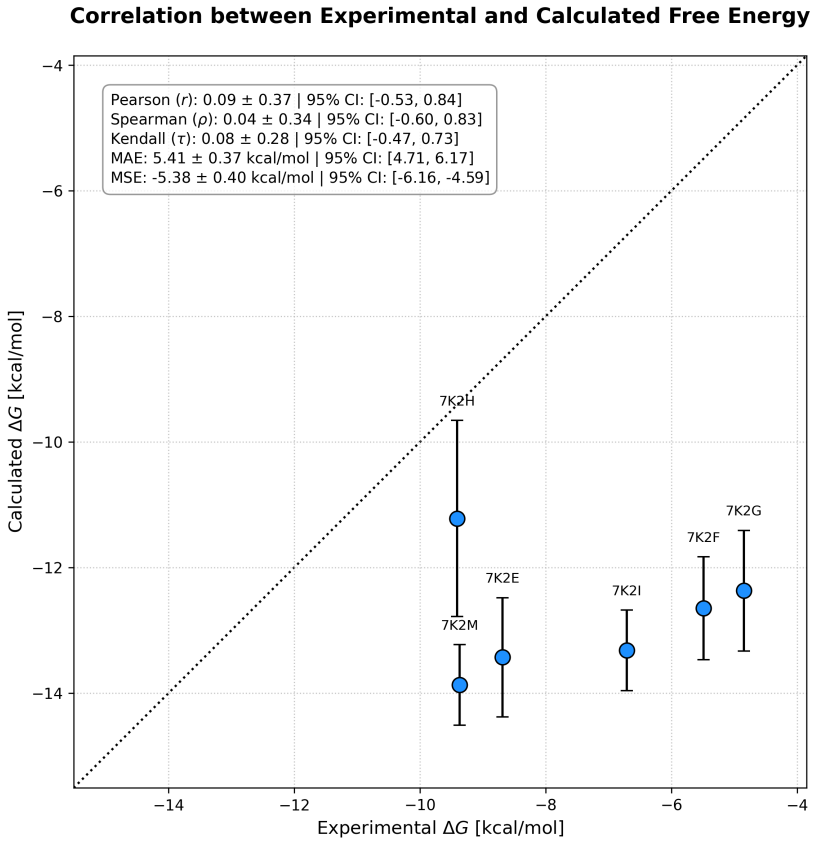
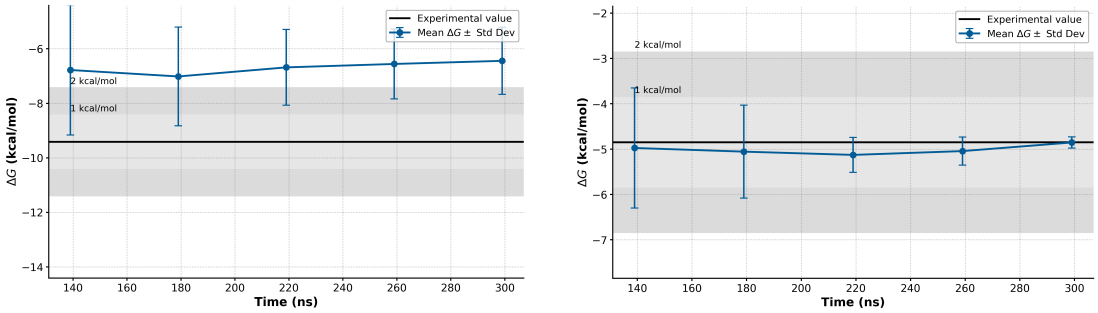


Figure 3.6: Correlation between simulated ΔG_b° and experimental values across the 7K2* systems. The dotted line is $y = x$. Correlation coefficients, mean absolute error (MAE) and mean signed error (MSE) are reported in the text box.

3.1.2 a99SB-disp: Under-binding for Rigid Complexes

Convergence and accuracy. Figure 3.7 shows the same analysis done in the previous section (figure 3.1), but now for a99SB-disp. For 7K2H, the standard deviation is always larger than 1 kcal/mol but it decreases from 2.4 kcal/mol to 1.2 kcal/mol going from 140 ns to 300 ns. We cannot say the profile has converged but the standard deviation across the three independent simulations decreases in each successive window. The mean estimate is around -6.5 kcal/mol, which is above the experimental value; this is the under-binding we expected for the a99SB-disp force field with relatively rigid peptides. For 7K2G the results are completely different from what we have seen up to now: the cross-replica standard deviation is well below 1 kcal/mol in the last three windows and this value seems to be steadily decreasing, reaching a standard deviation of 0.12 kcal/mol at 300 ns. Following what we have said in section 2.4.1, we cannot say that the profile has converged simply because of the length of the simulation. If, by extending the simulation, the standard deviation will still be below 1 kcal/mol, then we

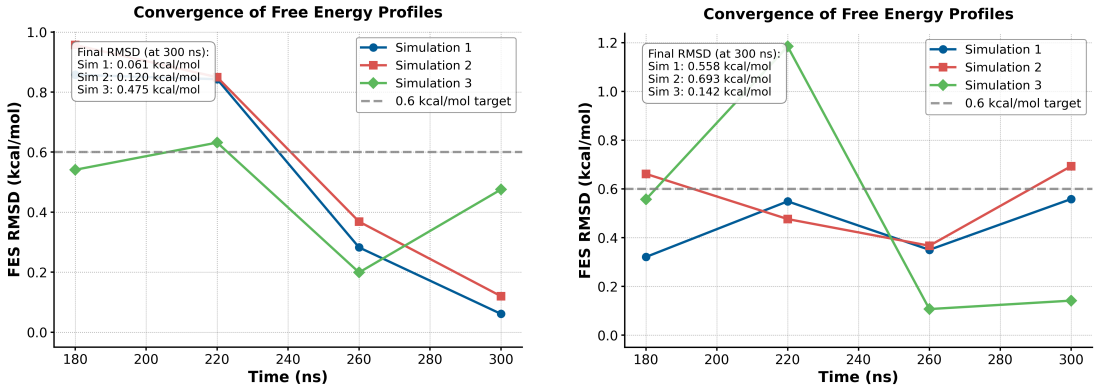
will say that the system has converged, at least according to the criterion presented in 2.4.1. The measure for this system is also accurate as the binding free energy is around -5 kcal/mol with an experimental binding free energy of -4.85 kcal/mol.



(a) 7K2H (b) 7K2G

Figure 3.7: Absolute binding free energy $\Delta G_b^\circ(t)$ (mean $\pm$ std over three independent simulations) for 7K2H and 7K2G with a99SB-disp. The black lines mark the experimental values.

The picture we get from the FES-RMSD analysis in figure 3.8 is a bit deeper. For 7K2G the FES-RMSD is below the target of 0.6 kcal/mol in the penultimate window but not in the last one, where both simulation 1 and 2 see an increase in their FES-RMSD value. The second convergence criterion is crucial, in this case, to understand that the system has definitely not converged yet, since as we can see also in figure 3.9b, the unbound state population is still evolving meaning that the bias is still changing and the FES-RMSD is still above target. For 7K2H we also have an interesting behaviour: simulations 1 and 2 follow almost the same path and their FES-RMSD decreases at each window, while simulation 3 is going up and down. Then, even if the FES-RMSD has been below the target of 0.6 kcal/mol for the last two windows, we cannot say that the system has definitely converged.

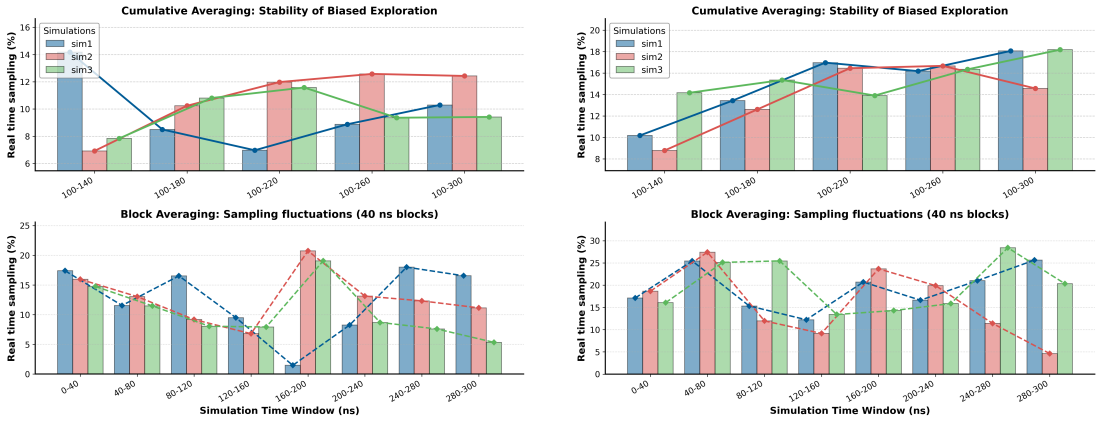


(a) 7K2H (b) 7K2G

Figure 3.8: FES-RMSD between successive 40 ns blocks for the three independent simulations of 7K2H and 7K2G with a99SB-disp. The dashed lines represent the 0.6 kcal/mol target.

Sampling efficiency. Figure 3.9 shows the real-time sampling of the unbound state described in section 2.4.2, both as a cumulative average and as a block average over 40

ns windows. Focusing on the sampling of the unbound state is even more important in this case, as we want to compare the unbound state population with what we got in the simulations with Amber14SB. For 7K2G, the three independent simulations show a consistent behaviour and, more importantly, they all have a substantially higher real-time unbound population (between 14% and 18%) compared to the simulations with Amber14SB (see figure 3.3b). This is the demonstration that a99SB-disp is able to overcome the over-compaction issue of Amber14SB, as the peptide is able to explore more extensively the unbound state. Moreover, the enhanced exploration directly translates in a better ΔG_b° estimate as we have seen in figure 3.7b. For 7K2H, the real-time unbound population ranges between 9% and 12%, showing that over-compaction is less strong for more rigid peptides and in fact the ΔG_b° estimate was quite accurate for 7K2H simulated with Amber14SB.



(a) 7K2H (b) 7K2G

Figure 3.9: Cumulative (top) and block-averaged (bottom) real-time sampling of the unbound state of three independent simulations of 7K2H and 7K2G with a99SB-disp.

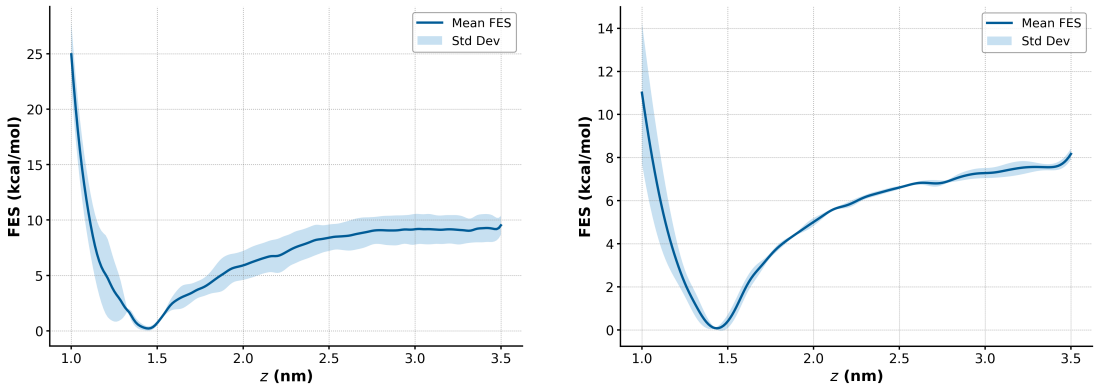
The global real-time populations for the bound, transition and unbound states, together with the total number of round trips are summarized in table 3.2 below. We can see that the number of round trips is higher for both systems with respect to their Amber14SB-simulated counterparts as the unbound state is sampled more extensively.

System	Sim	Bound (%)	Transition (%)	Unbound (%)	Round trips
7K2H	1	60.2	29.5	10.3	1209
	2	54.2	33.4	12.4	1367
	3	54.3	36.3	9.4	1044
7K2G	1	46.3	35.6	18.1	1738
	2	50.5	34.9	14.6	1651
	3	47.2	34.6	18.2	1864

Table 3.2: Real-time sampling percentages for the bound, transition and unbound states, together with the total number of round trips for systems 7K2H and 7K2G simulated with a99SB-disp.

In addition, we report in figure 3.10 the 1d FES along the funnel projection axis for 7K2H and 7K2G simulated with a99SB-disp. The comparison with respect to figure 3.4 is quite revealing, as we can see that the bound state is sampled quite well with both force fields and the main difference arises in the unbound state, where there is quite a

difference in the constant free energy value associated to the unbound state between figures 3.4 and 3.10.



(a) 7K2H (b) 7K2G

Figure 3.10: 1d Free Energy Surfaces (FES) along the funnel projection axis (z) for 7K2H and 7K2G with a99SB-disp. The deep free energy minimum associated to the bound state and plateau associated to the unbound state are still present but we can see the difference in the free energy value of the unbound state for both systems with respect to 3.4.

Finally, as for simulations with Amber14SB, exchange probabilities are stable between 30% and 60% for 7K2H and between 20% and 60% for 7K2G.

Correlation with peptide flexibility. Figure 3.11 shows the same correlation analysis presented in 3.5, but for a99SB-disp. As expected, we get a trend reversed in sign with Pearson $r = -0.94$. However this is not supported by the rank-based statistics: Spearman $\rho = -0.49$ and Kendall $\tau = -0.33$. They are far from statistical significance (since their p-value is well above 0.05) but more importantly their bootstrap confidence intervals span the entire $[-1, 1]$ range. This means that the Amber14SB over-compaction bias follows peptide flexibility systematically, while a99SB-disp underbinding does not and appears to be sensitive to sampling noise on this small dataset.

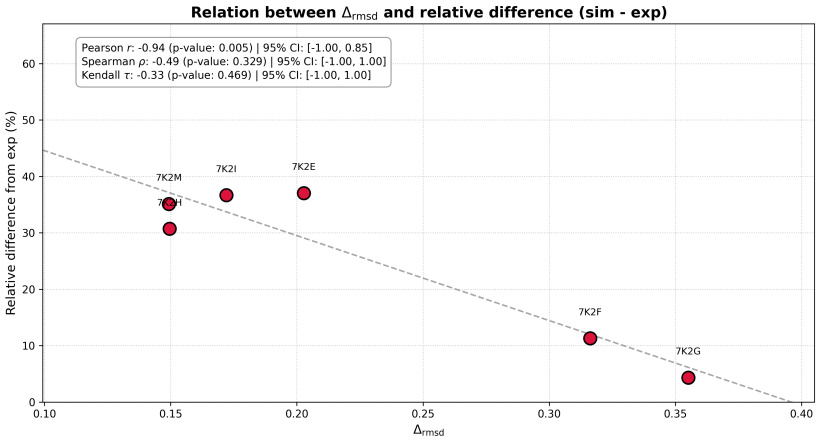


Figure 3.11: Relative error of the simulated ΔG_b° (with a99SB-disp) with respect to experimental values, as a function of Δ_{rmsd}, across all six 7K2* complexes.

Overall accuracy. Figure 3.12 shows the correlation between simulated ΔG_b° (with a99SB-disp) and its experimental value. All measures are better than what we got in 3.6 but still, the correlation is weak: Pearson $r = 0.47 \pm 0.28$, Spearman $\rho = 0.49 \pm 0.31$, Kendall $\tau = 0.37 \pm 0.25$. In particular the systematic bias introduced by a99SB-disp is more than half of the one introduced by Amber14SB: MAE $= 2.25 \pm 0.41$ kcal/mol and MSE $= +2.04 \pm 0.47$ kcal/mol. The positive bias of the MSE confirms the expected under-binding direction of the bias, which of course affects primarily the more rigid peptides, while actually obtaining very good results for more flexible peptides (7K2F and 7K2G). The correlation analyses with both force fields confirm that, while neither force field alone is accurate across the whole 7K2* family, the systematic error introduced by a99SB-disp on rigid complexes is substantially smaller than the one introduced by Amber14SB on flexible ones. This is an important result in light of a possible extension of the three-level screening protocol mentioned in chapter 4.

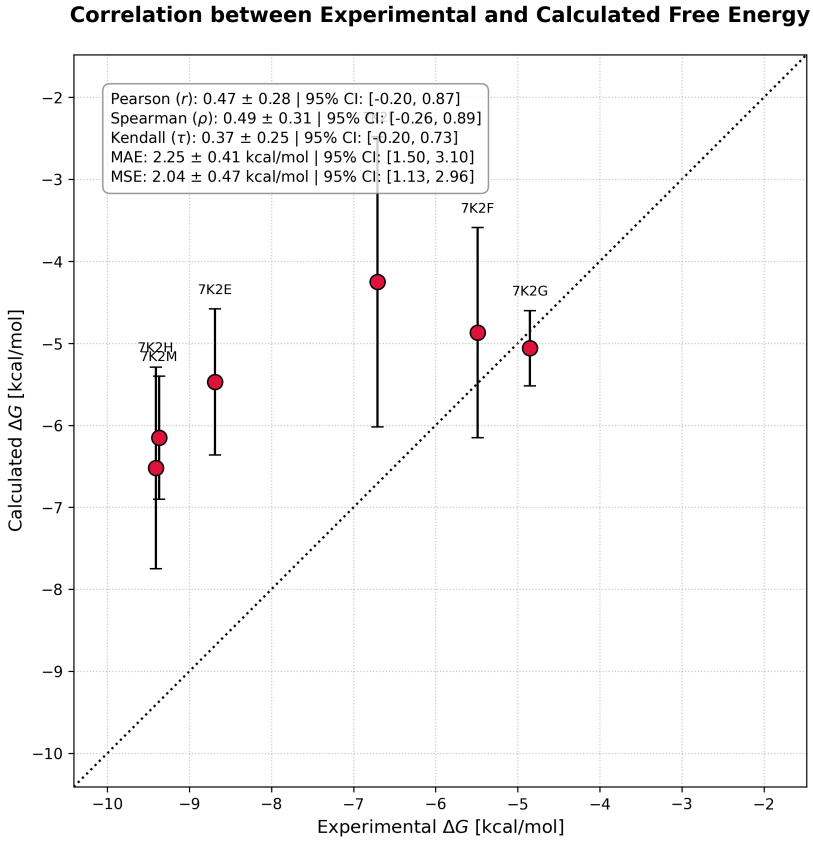


Figure 3.12: Correlation between simulated ΔG_b° (with a99SB-disp) and experimental values across all six 7K2* complexes. The dotted line is $y = x$.

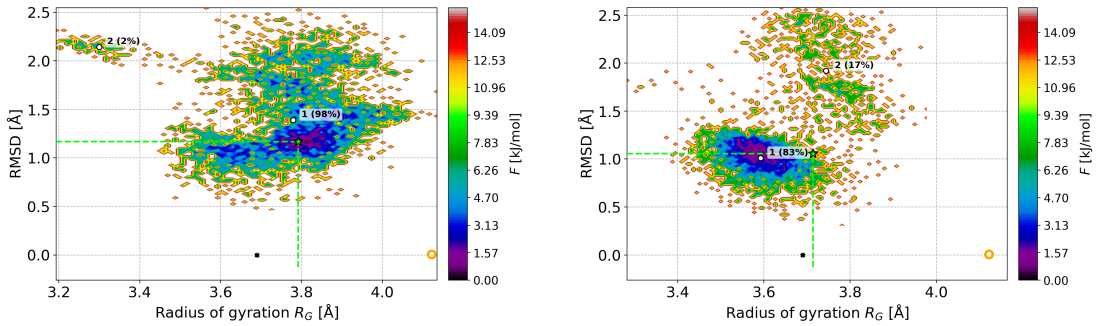
3.2 A Predictive Three-Level Screening Protocol

We design a three-level screening protocol that assigns, prior to any OneOPES/funnel metadynamics production run, the force field (or a combination of Amber14SB and a99SB-disp) expected to minimize the systematic bias for a given protein-cyclic peptide complex. The following classification is based exclusively on the intrinsic conformational flexibility of the *free peptide*, expressed through two complementary, dimen-

sionless quantities: the coefficient of variation of the backbone root mean-square deviation, $\Delta_{\text{rmsd}} = \sigma_{\text{rmsd}}/\mu_{\text{rmsd}}$, and the coefficient of variation of the radius of gyration, $\Delta_{R_G} = \sigma_{R_G}/\mu_{R_G}$. As detailed in section 3.2.2, Δ_{R_G} is used to separate structurally distinct complexes (Level 2), while Δ_{rmsd} is used to rank complexes belonging to the same family (Level 3). Both measures are intensive and dimensionless, allowing us to compare them across peptides of different sizes.

3.2.1 Conformational Sampling in Explicit and Implicit Solvent

Both quantitative Levels of the screening protocol (Level 2 and Level 3) use REMD simulations of the isolated peptide to extract information on its intrinsic flexibility, under two different solvation regimes. In explicit solvent, we run 50 ns REMD simulations with Amber14SB+TIP3P using 32 replicas, with temperatures in the range $300 - 450K$. In implicit solvent, we run 50 ns REMD simulations with Amber14SB using 8 replicas, with temperatures in the range $300 - 700K$. For each peptide, mean and standard deviation of R_G and rmsd are extracted directly from the obtained trajectories. Before presenting the three-level screening protocol, it is instructive to look at the conformational space explored by each peptide both in explicit and implicit solvent. Figure 3.13 shows the joint distribution of backbone rmsd and R_G sampled by 7K2G in both solvent regimes. The corresponding plots for the remaining peptides are reported in Appendix B.



(a) explicit (b) implicit

Figure 3.13: Two-dimensional free energy surface projected onto the backbone rmsd and radius of gyration, sampled by 7K2G during 50 ns REMD simulations both in explicit and implicit solvent. The black cross represents the reference structure for the rmsd.

The two plots reveal a slightly different conformational behaviour. In explicit solvent, the peptide populates a single, broad basin at $R_G \approx 3.7$ Å and which extends from 0.5 Å to 2.5 Å along the rmsd axis. A second, minor and more compact basin is present at low values of $R_G \approx 3.3$ Å and high values of rmsd. In implicit solvent, instead, we observe two highly populated basins, with the bigger one at $R_G \approx 3.5$ Å and rmsd ≈ 1.1 Å and the smaller one at $R_G \approx 3.75$ Å and higher values of rmsd between 1.5 Å and 2.5 Å. The main difference we can observe between the two plots is the fact that ensembles are continuously sampled in explicit solvent, while in implicit solvent there are several separate clusters. The broad and continuous sampling for explicit solvent turns out to be quite similar for all 7K2* complexes (see appendix B), which motivates the choice of Δ_{rmsd} just for Level 3, as explained in section 3.2.2.

3.2.2 Protocol details

In this section we introduce the details of the predictive screening protocol, followed by the resulting classification which will assign each complex to the respective force field (or force field combination). The protocol is composed of three levels.

Level 1: biological assessment. The first level is a qualitative analysis of the cyclic peptide sequence and its topology: size, type of cyclization (head-to-tail or backbone-to-side-chain), presence of additional constraints (such as internal disulphide bonds), possible additional information given by experimentalists who synthesized or discovered the peptide. This step does not play a crucial role, by itself, into assigning the peptide to a force field but it establishes to which structural family (that is, same protein target, same topology differing only by a small number of point substitutions) it belongs to and distinguishes peptides belonging to different families. For example, as we have seen in sections 2.2.1 and 2.2.2, 1SFI and the complexes 7K2* belong to different families and we expect the former to be more rigid than the latter. We also expect, following the observations made in the same sections, that peptides belonging to family 7K2* have different levels of flexibility. These distinctions (between different families and among peptides in the same family) are important to understand whether Level 2 alone is sufficient to rank a peptide based on its flexibility, or there's the need for Level 3, which is restricted to members of the same family.

Level 2: explicit-solvent REMD simulations. The second level introduces a quantitative analysis. Each peptide is simulated for 50 ns using REMD in explicit TIP3P solvent with Amber14SB. At this level, we compare complexes of different sizes and topology, as we consider the 7-residue 7K2* complexes and the 14-residue 1SFI: this is what motivates the choice of descriptor used for this level.

Δ_{rmsd} requires an optimal superposition to a bound-state reference conformation. However superposition-based rmsd is known to average out differences across regions with different similarity levels, making this measure not as informative when comparing structurally different conformational ensembles, which is why superposition-independent, distance-based measures have been proposed as an alternative [72]. Δ_{R_G}, instead, does not require a reference structure and it is just a geometric property of each individual conformation, which is directly comparable across different topologies and peptide sizes [73]. This is why we will use this measure for Level 2. Moreover, as shown in appendix E, using Δ_{rmsd} for Level 2 does not produce a biologically meaningful separation between complexes 1SFI and 7K2*, which is what we expect as rmsd depends on a bound-state reference structure, meaning it is more difficult to compare structurally different complexes.

In order to mathematically formalize the separation between flexibility regimes, we built a Kernel Density Estimate (KDE) [74] over the Δ_{R_G} values of all systems screened so far. KDE is one of the most well-known approaches to estimate the underlying probability density function of a dataset and it is a nonparametric density estimator requiring no assumption that the underlying density function is from a parametric family (e.g, Gaussian, Exponential, etc.). Given a sample of n independent observations $X_1, \ldots, X_n$, the KDE can be expressed as:

$$\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{3.1}$$

where K is a smooth function called Kernel function (typically Gaussian or Spherical), and $h > 0$ is the smoothing bandwidth that controls the amount of smoothing applied. In this case, we have chosen a Gaussian kernel, that is $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$, where $u = \frac{x - X_i}{h}$ and $h = \frac{1}{2}\sigma_{\text{data}}$. A more thorough theoretical background of KDE is provided in Appendix C.

Figure 3.14 shows the resulting KDE distribution as a function of Δ_{R_G} for all seven complexes considered in this thesis. We can see that the KDE is evidently bimodal, featuring a first, low-density mode around $\Delta_{R_G} \approx 0.02$ populated only by 1SFI, and a second, high-density mode around $\Delta_{R_G} \approx 0.030 - 0.039$ populated by all six 7K2* variants. The minimum of the KDE, between the two modes, which identifies the hard cut-off value is located at $\Delta_{R_G}^{\text{cut-off}} = 0.024$. 1SFI, that is well below the cut-off, is assigned directly to Amber14SB, consistent with the biological analysis done in section 2.2.2 and in Level 1 of the protocol. The whole 7K2* family is carried to Level 3, since their Δ_{R_G} are packed together; as expected, Level 2 is not able to resolve the difference in flexibility among peptides belonging to the same family (with identical topology and chain length) [73].

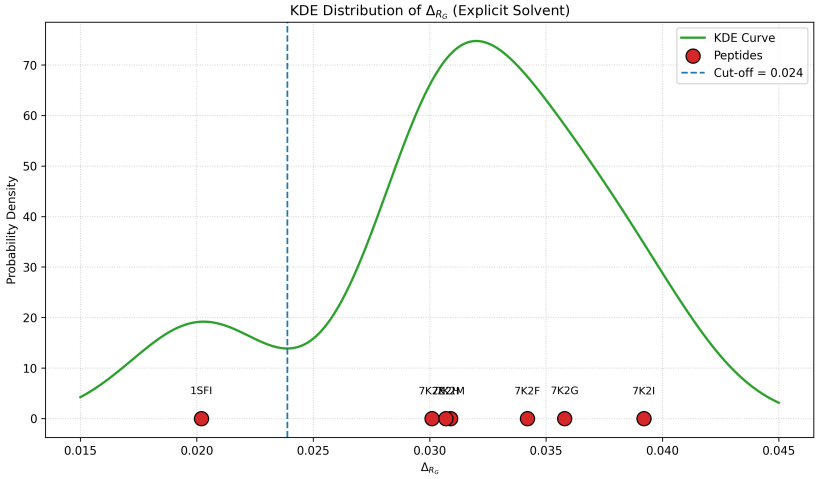


Figure 3.14: KDE distribution of Δ_{R_G} obtained from 50 ns explicit-solvent (Amber14SB+TIP3P) REMD simulations of the free peptides. 1SFI is below the data-driven cut-off at $\Delta_{R_G}^{\text{cut-off}} = 0.024$ and is assigned directly to Amber14SB, while 7K2* complexes are above the cut-off.

Level 3: implicit solvent REMD simulations. The third level of screening is applied only within a family for which Level 2 cannot discriminate between members, as in the case of the 7K2* family. As anticipated, here we use Δ_{rmsd} for two reasons: within a single structural family, each peptide shares the same topology and backbone length, so that the superposition required to compute RMSD is consistent across all peptides; moreover, unlike R_G, which is not sensitive to point substitutions that leave the majority of distances unchanged (as shown by Fuertes *et al.* in [73]), rmsd, since it is computed from a reference bound-state, is sensitive precisely to the local rearrangements caused by point substitutions. As shown in appendix E, using Δ_{R_G} for Level 3 (instead of Δ_{rmsd}) fails exactly because R_G barely changes across the point substitutions and the resulting KDE groups together five peptides in a single mode, isolating only 7K2G, which is in direct contradiction with what we expect following Level 1 (and

the analysis done with Δ_{rmsd} at Level 3). Each member of the 7K2* family is then simulated with a 50 ns implicit solvent REMD, as discussed in the previous section 3.2.1.

Applying the same KDE procedure (using a Gaussian Kernel and $h = \frac{1}{2}\sigma_{\text{data}}$) to the resulting Δ_{rmsd} values we get, again, a bimodal distribution (figure 3.15), with a first mode around $\Delta_{\text{rmsd}} \approx 0.15 - 0.20$ which groups together 7K2H, 7K2M, 7K2I and 7K2E and a second mode around $\Delta_{\text{rmsd}} \approx 0.31 - 0.36$, comprising 7K2F and 7K2G. The within-family cut-off is represented by the minimum of the KDE, at $\Delta_{\text{rmsd}}^{\text{cut-off}} = 0.267$: peptides falling to the left of the cut-off are classified as relatively ordered, and they are assigned to the hybrid approach (see next paragraph) combining Amber14SB and a99SB-disp, while peptides falling to the right of the cut-off are classified as disordered, and they are assigned to a99SB-disp. This is consistent with expectations built in section 2.2.1 on the basis of the E78P substitution and the loss of Asp77/Thr80 (respectively for 7K2F and 7K2G) side chains. A robustness assessment of this layer of the screening protocol is provided in Appendix D by performing Leave-One-Out cross-validation.

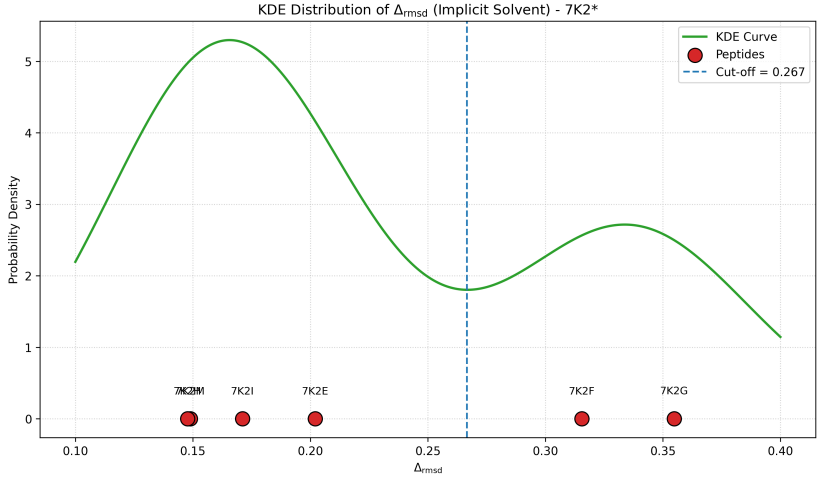


Figure 3.15: KDE distribution of Δ_{rmsd} from 50 ns REMD implicit solvent simulations of the free 7K2* peptides. The bimodal distribution separates an ordered subgroup (7K2E, 7K2H, 7K2I, 7K2M) from a disordered subgroup (7K2F, 7K2G) with a cut-off at $\Delta_{\text{rmsd}}^{\text{cut-off}} = 0.267$.

Then, the assignment obtained from the three-level screening protocol is the following: Amber14SB for 1SFI, hybrid approach for 7K2E, 7K2I, 7K2M, 7K2H and a99SB-disp for 7K2F and 7K2G as can be seen in the table 3.3 below. This assignment will be used to run multi-replica enhanced sampling simulations of the entire complexes (protein and cyclic peptide) with the respective force fields and results are presented in section 3.3.

The Hybrid approach: A consensus estimator. In order to study peptides which exhibit intermediate intrinsic flexibility, we introduce a *Hybrid Approach*. We know that the systematic biases coming from Amber14SB and a99SB-disp point in opposite directions; then, the hybrid approach can act as a consensus estimator. We perform two independent sets of multi-replica simulations for the same complex: one with Amber14SB and the other one with a99SB-disp and take the arithmetic mean of their

System	Amber14SB	Hybrid Approach	a99SB-disp
1SFI	X		
7K2E		X	
7K2F			X
7K2G			X
7K2H		X	
7K2I		X	
7K2M		X	

Table 3.3: Force field assignment for the 1SFI complex and the 7K2* family based on the three-level screening protocol.

absolute binding free energies:

$$\Delta G^{\circ}_{\text{hybrid}} = \frac{\Delta G^{\circ}_{\text{a14}} + \Delta G^{\circ}_{\text{a99}}}{2} \tag{3.2}$$

Note that this is an empirical correction applied to the final observable ΔG°_{b} and not a thermodynamic mixing of Hamiltonians or partition functions. Then, while the scalar ΔG°_{b} can be averaged to cancel opposite systematic biases of different force fields, the underlying one-dimensional free energy surfaces (FES) cannot be averaged as that would violate Boltzmann statistics. This approach is used here as a robust estimator for binding free energies to be compared with experimental data, but it cannot be used to study the binding-unbinding pathways for the systems we applied the hybrid approach on. This last point motivates the development of advanced sampling techniques capable of bypassing the force field dichotomy (whose very first application is presented in section 3.5).

Limitations and extension to new data. The main, clear limitation of the KDE-based cut-off procedure is the limited sample size (seven systems in total) which make both procedure (Level 2 and 3) extremely sensitive to the addition of further data points. Then, nothing guarantees us that the bimodality shown in Figures 3.14 and 3.15 will survive as the dataset grows. We assessed this point by doing a Monte Carlo stress test, where a certain number of synthetic points, taken from a uniform distribution spanning the Δ_{R_G} range for Level 2 and the Δ_{rmsd} range for Level 3, was added to the dataset. Clearly, peptide flexibility is not uniformly distributed, but this test gives us an idea of what happens to the bimodal distributions as we add more data points. The KDE was then recomputed and we recorded the fraction of trials where the distribution remained bimodal; in those cases also the cut-off value was reported. Note that we track the % of bimodality by tracking the number of modes of the distribution for each trial: if the number of modes (i.e., the number of maxima) is exactly two then we say that the distribution is bimodal. The results are summarized in the tables 3.4 and 3.5 below.

The results are quite different depending on the Level. Level 3 is rather robust, with 85.8% of the trials still bimodal after adding 10 extra synthetic points. On the other hand, Level 2 degrades pretty fast, reaching a mean fraction of bimodality around 46.6% after adding 10 extra synthetic points. This tells us that an extension of this protocol is probably needed as we extend the dataset. It should be noted, then, that the classification works well in this case (since, in particular, 1SFI is highly rigid) but it looks difficult to generalize as the dataset gets extended, especially for Level 2. Possible extensions of this protocol are discussed in chapter 4.

n_{extra}	% bimodal	cut-off mean	cut-off std
1	88.5%	0.0237	0.0006
2	77.7%	0.0236	0.0008
3	70.4%	0.0237	0.0012
5	59.7%	0.0239	0.0020
8	51.8%	0.0241	0.0026
10	46.6%	0.0245	0.0029

Table 3.4: Bimodality stress test for Level 2 (explicit solvent, Δ_{R_G}). n_{extra} synthetic points, drawn uniformly from $[0.015, 0.045]$, are added to the real dataset ($n = 7$, original cut-off 0.024) over 1000 Monte Carlo trials per n_{extra}; the table reports the fraction of trials in which the resulting distribution remained bimodal, and the mean and standard deviation of the resulting cut-off across the bimodal trials.

n_{extra}	% bimodal	cut-off mean	cut-off std
1	100.0%	0.267	0.010
2	97.7%	0.267	0.013
3	93.4%	0.264	0.015
5	91.3%	0.263	0.019
8	86.1%	0.260	0.023
10	85.8%	0.262	0.025

Table 3.5: Bimodality stress test for Level 3 (implicit solvent). n_{extra} synthetic points, drawn uniformly from $[0.10, 0.40]$, are added to the real dataset ($n = 6$, original cut-off 0.267) over 1000 Monte Carlo trials per n_{extra}; the table reports the fraction of trials in which the resulting distribution remained bimodal, and the mean and standard deviation of the resulting cut-off across the bimodal trials.

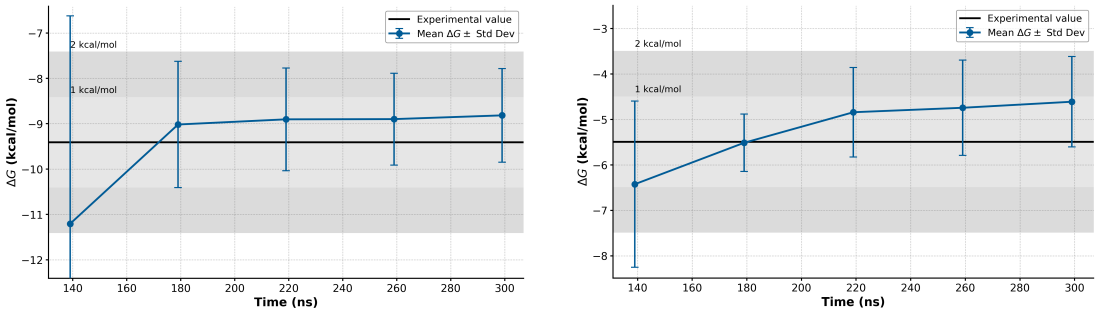
3.3 Application of the Predictive Screening Protocol

We present now the results obtained by applying the screening protocol (see section 3.2) that assigns each of the 7K2* complexes the force field (or their combination) in order to get an estimate of the binding free energy, ΔG_b°. We recall the results of the assignment, which are also summarized in table 3.3: 1SFI is assigned to Amber14SB directly at Level 2; within the 7K2* family, Level 3 classifies 7K2E, 7K2H, 7K2I, 7K2M as relatively ordered and assigns them to the hybrid approach, while 7K2F and 7K2G are classified as disordered and assigned to a99SB-disp. The results for 7K2G, simulated with a99SB-disp have already been presented in the previous section, so here we use two different representatives: 7K2H, to show a hybrid approach assignment, and 7K2F to show a second a99SB-disp assignment for which no analysis has been presented. The results for the remaining systems (7K2E, 7K2I, 7K2M) are reported in Appendix G.3.

Multi-run convergence. Figure 3.16 below shows the $\Delta G_b^\circ(t)$ analysis presented in section 2.4.1 for 7K2H (with the hybrid approach) and 7K2F (with a99SB-disp). For the hybrid approach, the error bars are obtained by propagating the cross-replica standard deviations of the Amber14SB and a99SB-disp estimates: $\sigma_{\text{hybrid}} = \frac{1}{2}\sqrt{\sigma_{\text{a14}}^2 + \sigma_{\text{a99}}^2}$. Moreover, as explained in section 3.2, since we have not considered a "hybrid" Hamiltonian but simply the average of the ΔG_b° estimates, then the analysis in figure 3.16 is the only convergence check available for hybrid systems as the FES-RMSD and population

analyses cannot be defined for hybrid systems.

Looking at figure 3.16, for 7K2H the standard deviation decreases from 4.6 kcal/mol at 140 ns to 1.03 kcal/mol at 300 ns, remaining above the 1 kcal/mol threshold for the whole $\Delta G_b^\circ(t)$ curve meaning the profile has not converged yet. More importantly, the mean hybrid estimate is around -8.8 kcal/mol in the last window, which is rather close to the experimental value of -9.41 kcal/mol. The better estimate for 7K2H (with respect to both isolated force fields) is a direct consequence of the error cancellation the hybrid approach is developed to produce. For 7K2F, which was simulated with a99SB-disp, the standard deviation is below 1 kcal/mol from the second window (180 ns). However we can see that it is still changing (it increases from the second to third window), indicating that the three simulations are still trying to agree on which conformations have more weight; this is also confirmed by the block analysis in figure 3.18. Moreover, the mean estimate which is around -4.7 kcal/mol and pretty close to the experimental value of -5.49 kcal/mol, presents a slight drift toward higher (less negative) values of ΔG_b° which should not happen if the system were truly converged.



(a) 7K2H (hybrid approach)

(b) 7K2F (a99SB-disp)

Figure 3.16: Absolute binding free energy $\Delta G_b^\circ(t)$ evaluated at five sparse time points following section 2.4.1, for 7K2H (a, hybrid approach) and 7K2F (b, a99SB-disp). Black lines mark the experimental values.

FES-RMSD and sampling efficiency. Since, as we have said, the second convergence check can only be applied to systems simulated with a single force field, we restrict this analysis to 7K2F. Figure 3.17 shows the FES-RMSD between successive 40 ns windows for three independent simulations. All three simulations reach a final FES-RMSD value below the target of 0.6 kcal/mol but as we have said in section 2.4.2, it is not sufficient that a single window is below the target for all three simulations to claim convergence: it should be below it for several windows. What we can note, however, is the downward trend, as the FES-RMSD decreases in each successive window. Figure 3.18 shows the real-time sampling of the unbound state, both as a cumulative average and as a block average over 40 ns windows. The cumulative average between the three simulations is consistently between 11% and 13% and considerably higher than the sampling obtained with Amber14SB (see figure G.3). We can also note that simulation 3 is less stable (see block averaging) but it recovers toward the end of the simulation. This characteristic is also reflected in the FES-RMSD plot in figure 3.17.

The global real-time populations and total number of round trips are summarized in table 3.6 below. The high number of round trips confirms the excellent sampling efficiency.

Figure 3.19 shows the 1d FES along the funnel projection axis for 7K2F with a99SB-

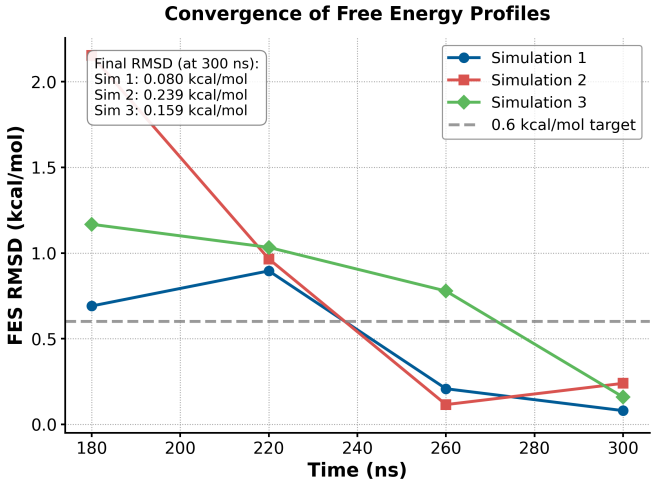


Figure 3.17: FES-RMSD between successive 40 ns blocks for three independent simulations of 7K2F with a99SB-disp. The dashed line marks the 0.6 kcal/mol target.

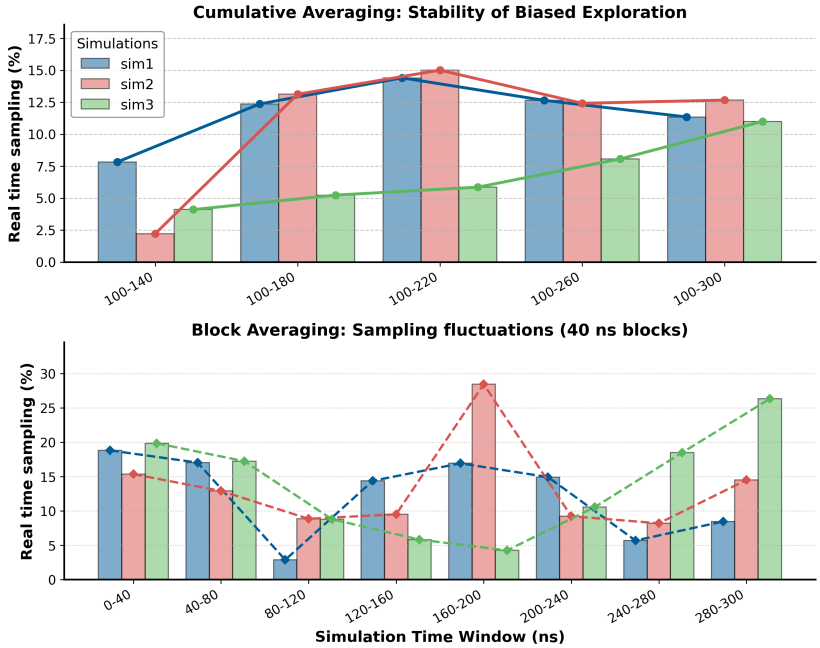


Figure 3.18: Cumulative (top) and block averaged (bottom) real-time sampling of the unbound state for three independent simulations of 7K2F with a99SB-disp.

System	Sim	Bound (%)	Transition (%)	Unbound (%)	Round trips
	1	55.4	33.2	11.3	1245
7K2F	2	49.5	37.8	12.7	1294
	3	57.1	31.9	11.0	1114

Table 3.6: Real-time sampling percentages for the bound, transition and unbound states, together with the total number of round trips for 7K2F simulated with a99SB-disp.

disp. The bound state appears slightly shifted towards higher values of z with respect to simulations with Amber14SB (see figure G.4) and the unbound state has a lower free energy value, correctly solving the over-compaction produced by Amber14SB.

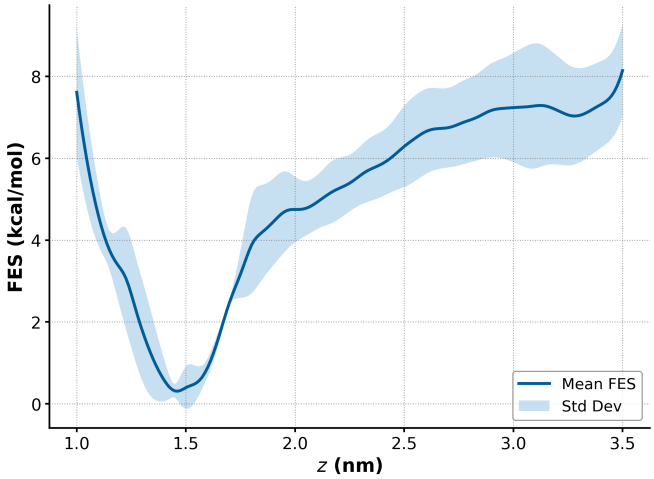


Figure 3.19: 1d Free Energy Surface (FES) along the funnel projection axis (z) for 7K2F simulated with a99SB-disp. The FES at higher values of z is not completely flat indicating that weights in the unbound state have not stabilized yet.

Finally, the exchange rate for 7K2F is between 20% and 60%, following the same pattern we presented previously: higher acceptance rate for low-index replicas, lower acceptance rate for high-index replicas.

Overall protocol accuracy. Figure 3.20 shows the correlation between the simulated ΔG_b° and the experimental values, where the simulated values have been obtained following the screening protocol prediction. As we did for the single force field correlations, all statistical measures were not computed as single point estimates but using a Monte Carlo error propagation, as explained in section 3.1.1.

The improvement over either single force field analysis, presented in section 3.1.1 and 3.1.2, is clear and substantial. The correlation measures are stronger: Pearson $r = 0.85 \pm 0.09$, Spearman $\rho = 0.70 \pm 0.16$ and Kendall $\tau = 0.55 \pm 0.20$. The systematic bias is also reduced: MSE $= -0.44 \pm 0.36$ kcal/mol which indicates still a slight tendency to over-binding, and the resulting MAE $= 1.04 \pm 0.27$ kcal/mol falls within the 1 kcal/mol accuracy commonly used as the target for a computational method to be useful in distinguishing binders of different affinity. This result is the main quantitative justification for the three-level screening developed in this thesis.

3.4 1SFI: an Out-of-Family Rigid Complex

We present now 1SFI, the only complex in this thesis that does not belong to the 7K2* family. As seen in section 2.2.2, 1SFI is structurally different with respect to the 7K2* peptides, since it is a 14-residue macrocycle (twice the size of the 7K2* complexes), contains an internal disulphide bond and it binds to a different protein target. It was chosen precisely because of these differences: it is larger, more rigid and has a more extensive protein-peptide interface than any 7K2* complex, making it, in theory, a considerably more difficult test for the multi-replica protocol presented in section 2.3;

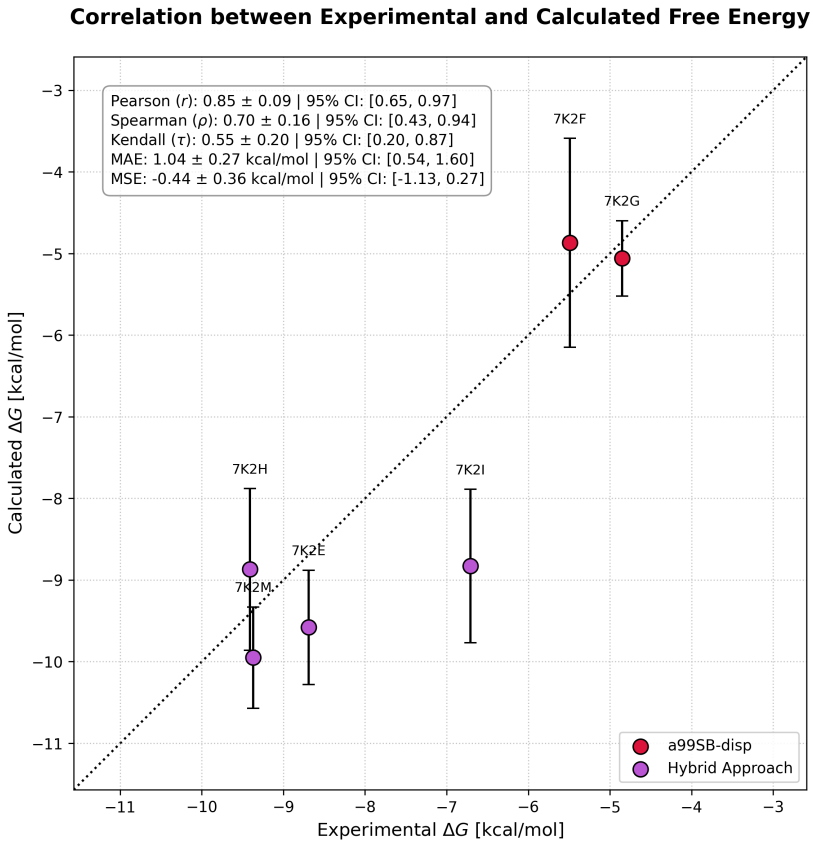


Figure 3.20: Correlation between simulated (using the screening protocol classification) and experimental ΔG_b° across all six 7K2* complexes. Purple points are assigned to the hybrid approach (7K2E, 7K2H, 7K2I, 7K2M); red points are assigned to a99SB-disp (7K2F, 7K2G). The dotted line is $y = x$. Correlation coefficients, MAE and MSE are reported in the text box with their corresponding 95% confidence intervals.

it is interesting to see if we can still correctly describe the binding-unbinding events and produce an accurate estimate of the binding free energy. Moreover, 1SFI is the only complex in Level 2 of the three-level screening protocol (3.2) that is not part of the 7K2* family and it is 1SFI's Δ_{R_G} value that allows the KDE to produce a bimodal separation between different structural families (see 3.14).

Because of its different size, we consider three independent simulations of 500 ns, discarding the first 200 ns as equilibration; the windows used for the ΔG_b° analysis, the block population analysis and the FES-RMSD are 50 ns. Following the assignment from Level 2 of the screening protocol (see table 3.3), 1SFI is classified as highly rigid and it is simulated with Amber14SB.

Convergence and accuracy. Figure 3.21 shows the $\Delta G_b^\circ(t)$ analysis for 1SFI simulated with Amber14SB, evaluated at six points every 50 ns after equilibration. The cross-replica standard deviation is below 1 kcal/mol until 400 ns, decreasing from 0.84 kcal/mol (at 250 ns) to 0.51 kcal/mol at 400 ns, but then it increases (together with its ΔG_b° estimate) over the last 100 ns, reaching a value of 1.30 kcal/mol in the last window. This behaviour is useful, as it shows that the apparent convergence was only a transient and not a true stabilization: this is why the cross-replica standard

deviation needs to be below the target for several windows since, as we can see, being below the target in one window is not a proof of convergence.

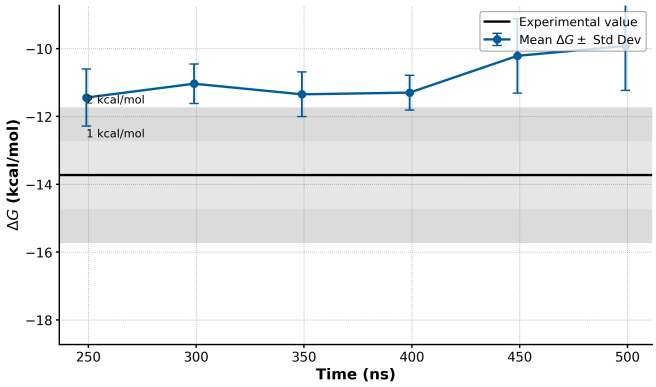


Figure 3.21: Absolute binding free energy $\Delta G_b^\circ(t)$ (mean $\pm$ std over three independent simulations) for 1SFI with Amber14SB, evaluated at six time points every 50 ns. The black line marks the experimental value.

The FES-RMSD in figure 3.22 confirms what we have said for figure 3.21. We can see that only in the last window the FES-RMSD drops below the 0.6 kcal/mol target value for all three simulations. Moreover, simulations 1 and 2 are still heavily fluctuating as is confirmed also by the block averaging in figure 3.23.

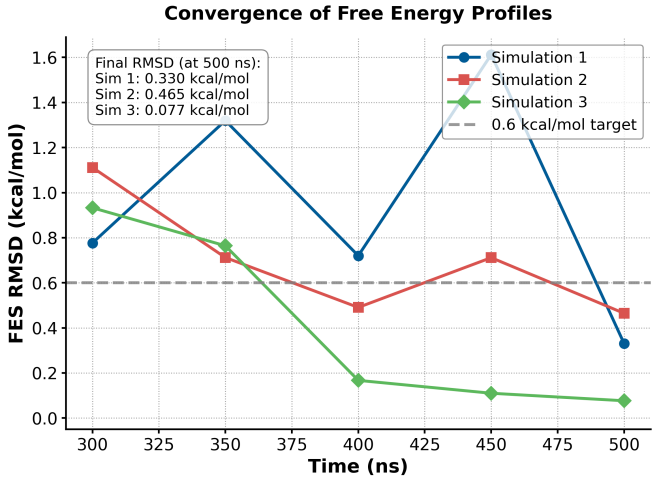


Figure 3.22: FES-RMSD between successive 50 ns blocks for three independent simulations of 1SFI with Amber14SB. The dashed line marks the 0.6 kcal/mol target.

Sampling efficiency. Figure 3.23 shows the real-time sampling of the unbound state both as a cumulative average and as a block average over 50 ns windows. As we did for the 7K2* complexes, also here the difference between bound, transition and unbound state is based only on `pp.proj`: the native bound pose is at `pp.proj` $\approx$ 2.1 nm, if `pp.proj` $<$ 2.4 nm the peptide is in the bound state, if `pp.proj` $>$ 3.8 nm the peptide is in the unbound state, if it is in between these values it is in the transition state.

As expected for a highly stable and rigid peptide the unbound state sampling is much lower with respect to what we get for the 7K2* complexes, with the unbound

state population ranging between 3% and 5%. Moreover, timescales for this system seem to be much longer than 7K2*, as we can see in the block averaging below: there are entire windows where the unbound state exploration is near 0% and then it goes up to higher values. This suggests that probably longer windows and longer simulation times are needed for this system to accurately explore its conformational space and get a better estimate of ΔG_b°.

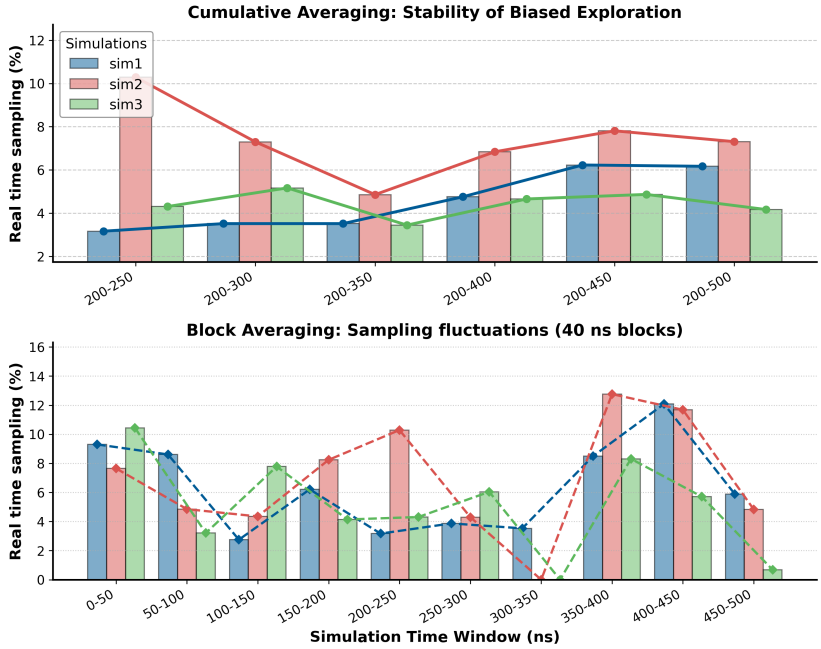


Figure 3.23: Cumulative (top) and block averaged (bottom) real-time sampling of the unbound state for three independent simulations of 1SFI with Amber14SB.

The global real-time bound, transition and unbound populations, together with the total number of round trips are summarized in table 3.7 below. Despite the low unbound population, the number of round trips confirms (together with the exchange probabilities) that the system is not kinetically trapped.

System	Sim	Bound (%)	Transition (%)	Unbound (%)	Round trips
1SFI	1	71.1	24.1	4.8	1126
	2	70.3	25.1	4.6	1268
	3	75.0	21.8	3.3	797

Table 3.7: Real-time sampling percentages for the bound, transition and unbound states, together with the total number of round trips for 1SFI with Amber14SB.

In addition, figure 3.24 shows the 1d FES along the funnel projection axis for 1SFI. Note that the spatial parameters of the funnel for 1SFI differ from the ones used for 7K2*, as the systems have different sizes: in this case, the funnel is longer and the upper restraining wall is positioned at $z = 4.5$ nm instead of $z = 3.3$ nm. The profile exhibits a deep energy minimum corresponding to the bound state, which is perfectly sampled but it also appears a second metastable state around $z \approx 3.0$ nm, which could be caused by the larger size of the peptide, representing a conformation where the peptide is trying to find the proper native contacts to fully re-enter the binding pocket. Finally, the flat region representing the unbound state is reached starting from $z = 3.5$ nm.

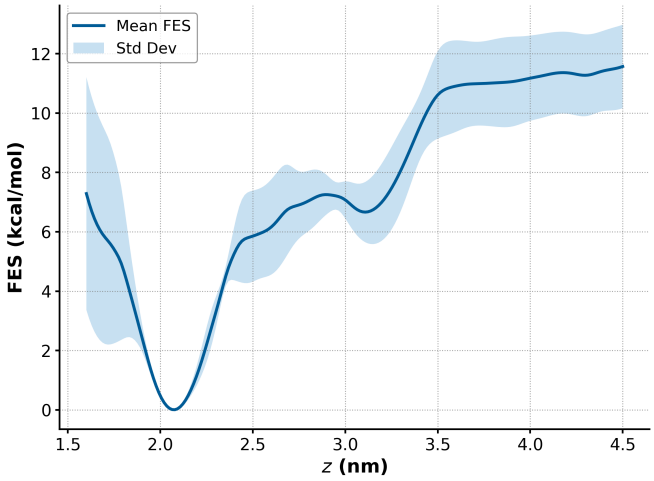


Figure 3.24: 1d Free Energy Surface (FES) along the funnel projection axis (z) for 1SFI simulated with Amber14SB.

Replica-exchange statistics are once again great and they follow the pattern of previous systems with exchange rates between 20% and 60% and higher exchange rates in lower-index replicas and lower exchange rates in higher-index replicas.

Overall correlation. In figure 3.25 we add the 1SFI ΔG_b° estimate in the correlation plot presented in figure 3.20, which now includes all seven complexes studied in this thesis. The Pearson correlation coefficient decreases, from $r = 0.85 \pm 0.09$ to $r = 0.77 \pm 0.11$ since 1SFI's linear distance from the $y = x$ line is the largest in the whole dataset. The rank-based statistics slightly improve as the simulation correctly predicts 1SFI to be the strongest binder: Spearman ρ increases from 0.70 ± 0.16 to 0.72 ± 0.16 and Kendall τ from 0.55 ± 0.20 to 0.58 ± 0.19. As expected, the systematic error metrics worsen: MAE $= 1.38 \pm 0.29$ kcal/mol and most notably MSE $= +0.12 \pm 0.34$ kcal/mol, which is caused by the under-binding observed for 1SFI. This is somewhat surprising considering the over-binding observed for the 7K2* complexes in figure 3.6 and it could be explained both by the lack of convergence of 1SFI and by the fact that this complex is inherently rigid, meaning Amber14SB over-binding plays less of a factor for this system, as we have also seen for 7K2H in section 3.1.1.

3.5 Beyond Heuristics: Preliminary Results of the Advanced Protocol

The three-level screening protocol presented in sections 3.2 and 3.3 relies on the choice of two fixed force fields (or their arithmetic average) and as we already discussed in section 3.2, this choice has an intrinsic limitation: through the hybrid approach we can correct opposite biases coming from Amber14SB and a99SB-disp on ΔG_b° but we cannot reconstruct the underlying free energy surface since the two force fields describe different unbound state ensembles. Here we present preliminary results of an alternative strategy, whose aim is to overcome the force field dichotomy by acting directly on the multi-replica simulation setup.

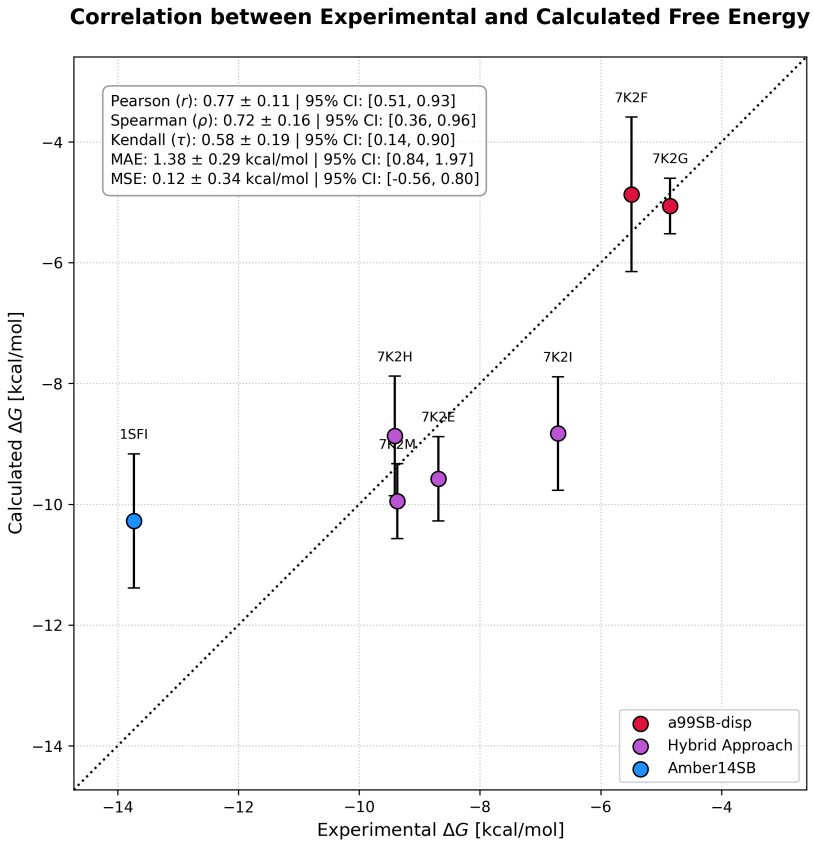


Figure 3.25: Correlation between simulated and experimental ΔG_b°, including 1SFI (blue), together with the hybrid approach (purple) and a99SB-disp (red) 7K2* predictions. The dotted line is $y = x$. Correlation coefficients, MAE and MSE with the respective 95% confidence intervals are reported in the text box.

Setup. The first seven replicas (indices 0–6) follow the same architecture described in the standard multi-replica protocol (section 2.3): each replica contains the entire protein-peptide complex, the same collective variables (`pp.proj`, `cmap` and hydration CVs) are biased and OneOPES parameters are identical to the ones already used up to now in this thesis. The only difference with respect to table 2.2 is that the addition of auxiliary hydration CVs is now no longer linear: we introduce a jump of two CVs between replica 1 and replica 2, to compensate for the fact that replica 7 does not contain the standard CVs anymore but it is used specifically for the isolated peptide.

Replica 7 contains only the free cyclic peptide, isolated from the protein. However, since in GROMACS all replicas must have the same number of atoms, we considered the following approach:

- We took a snapshot extracted from a previous simulation with Amber14SB where the peptide was already in the unbound state, precisely at `pp.proj` ≈ 2.2 nm (the bound state is at `pp.proj` ≈ 1.47 nm).
- A lower restraining wall is placed at `pp.proj` ≈ 2.1 nm to keep the peptide in the unbound region for the whole simulation, so that it does not interact with the protein.
- We keep replica 7 at $T = 350K$, the same temperature of replica 6 (to try to

maximize the potential energy overlap between replicas) but, while replica 6 biases `pp.proj`, `cmap` and hydration CVs, replica 7 biases two purely conformational collective variables: two $C_\alpha - C_\alpha$ distances, taken between residues at opposite sides of the peptide in order to form a cross pattern, so that when one distance elongates the other one contracts and vice versa. These CVs are designed to enhance the exploration of the peptide in the unbound state to target the over-compaction issue produced by Amber14SB. Then, through replica-exchange new configurations explored by replica 7 should be transferred to lower-index replicas.

The preliminary protocol was tested on 7K2G, as it is the system for which Amber14SB showed the largest error with respect to experiments. Three independent simulations of 100 ns were performed, way shorter than the 300 ns considered for the results presented in sections 3.1–3.4. Because of this, the analyses presented for other systems are not presented in the same form here: the equilibration cut-off and the window sizes used for both the $\Delta G_b^\circ(t)$ and FES-RMSD are reduced to 30 ns and 10 ns, respectively. This is the reason why the results obtained here should be seen as a preliminary and qualitative assessment of the strengths and limitations of this protocol.

Convergence. Figure 3.26 shows the $\Delta G_b^\circ(t)$ analysis at seven points, every 10 ns after 30 ns of equilibration. The cross-replica standard deviation remains around 0.9 kcal/mol, while the mean-estimate is stable around -10.7 kcal/mol. This is a slight improvement over the Amber14SB estimate for 7K2G presented in section 3.1.1 (which was around -12.5 kcal/mol), even though it is still pretty distant from both the estimate obtained with a99SB-disp and the experimental value, at -4.85 kcal/mol.

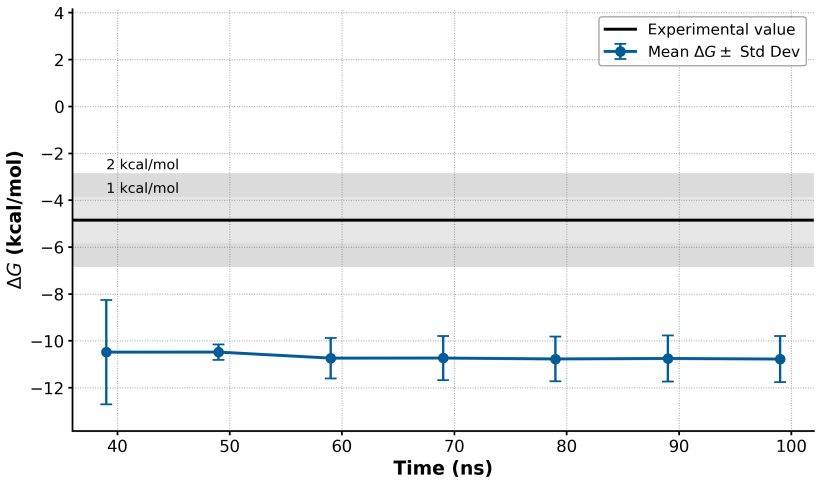


Figure 3.26: Absolute binding free energy $\Delta G_b^\circ(t)$ for 7K2G, shown as the seven-point analysis with 10 ns windows. The black line marks the experimental value.

Figure 3.27 shows the FES-RMSD between successive 10 ns windows. We can see that the curves are rather stable for all three simulations and that in the last window all three points are below the 0.6 kcal/mol target. Note, however, that this is most likely an artifact generated by the fact that we are using 10 ns windows, which are probably too short to capture slow conformational rearrangements that the CVs of replica 7 are designed to enhance. In the same way for which a single window below target in the longer $300 - 500$ ns simulations was not sufficient to establish convergence, here the same point holds. Despite this observations, the fact that errors for both analyses

remain below targets is a modestly encouraging sign for the short-term stability of the reweighting procedure.

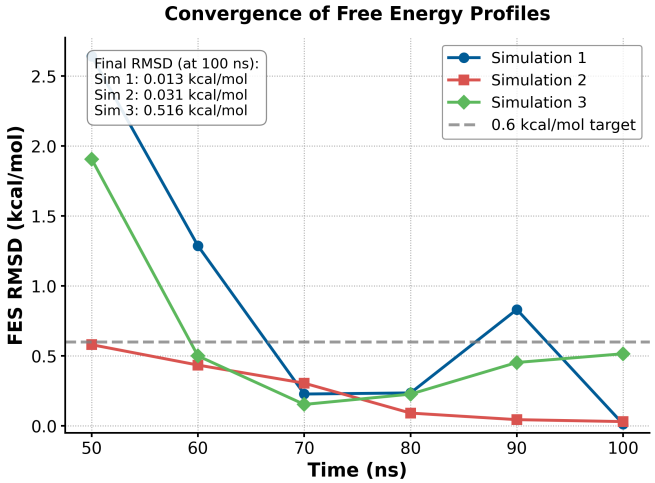


Figure 3.27: FES-RMSD between 10 ns blocks for three independent simulations of 7K2G. The dashed line marks the 0.6 kcal/mol target. Given the short block length, apparent stability should not be read as convergence (see main text).

Sampling efficiency. Figure 3.28 shows the real-time sampling of the unbound state both as a cumulative average and as a block average of 10 ns windows. We can see that the unbound state population differs heavily between different simulations: 5.0%, 8.5%, 11.4%. This variance directly correlates with the exchange rate between replicas 6 and 7 (between the entire complex and the isolated peptide), which is really low (0.66%, 0.98% and 1.29%, respectively). This difference in exploration of the conformational space could be due to the length of the simulation (at 100 ns, which is very short) and to the quite different exchange rate between independent simulations.

The global real-time bound, transition and unbound populations, together with the total number of round trips are summarized in table 3.8 below.

System	Sim	Bound (%)	Transition (%)	Unbound (%)	Round trips
	1	64.4	30.6	5.0	241
7K2G	2	56.7	34.8	8.5	352
	3	52.4	36.2	11.4	417

Table 3.8: Real-time sampling percentages for the bound, transition and unbound states, together with the total number of round trips for 7K2G simulated with the advanced multi-replica protocol.

We report in figure 3.29 the 1d FES along the funnel projection axis. The most noticeable difference compared to the standard Amber14SB simulation of 7K2G (see figure 3.4b) is in the unbound state (for $z > 2.5$ nm) as the free energy plateau is lower than the standard Amber14SB estimate. This also reflects the partial mitigation of the over-compaction artifact, which in turn explains the slightly more accurate binding free energy obtained in figure 3.26.

Note that the exchange rates between replicas 6 and 7 (0.6%-1.3%) are an order of magnitude lower than the ones observed for all other systems in this thesis (see sections

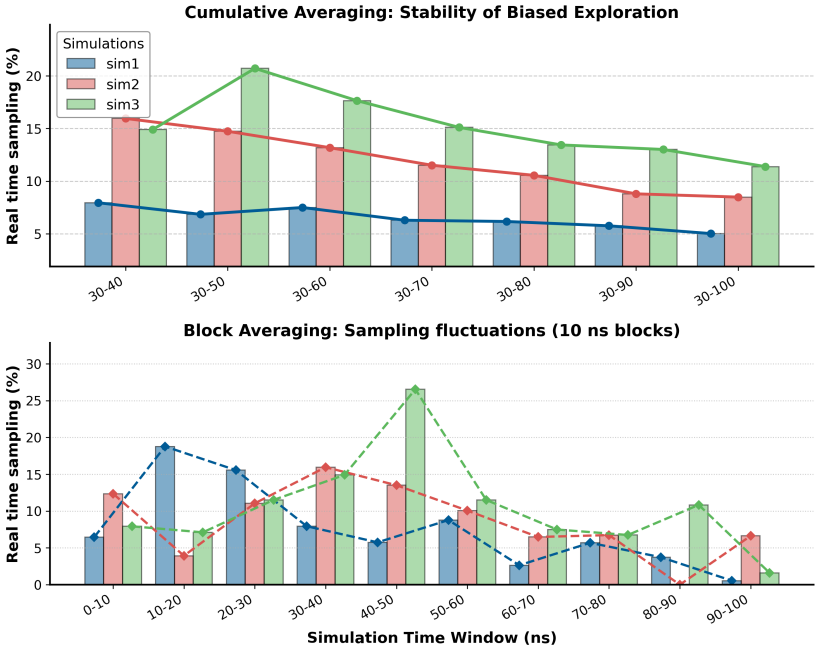


Figure 3.28: Cumulative (top) and block averaged (bottom) real-time sampling of the unbound state for three independent simulations of 7K2G with the advanced protocol.

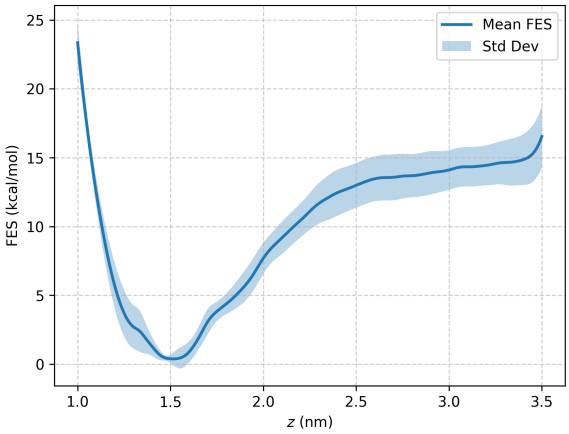


Figure 3.29: 1d Free Energy Surface (FES) along the funnel projection for 7K2G simulated with the advanced protocol. The unbound state plateau is lower compared to the standard Amber14SB simulation.

3.1.1, 3.1.2, 3.4). This remains the central limitation of the protocol as currently implemented: replica 6 and replica 7 have a completely different set of CVs, which limits the potential energy overlap and impacts the ΔG_b° estimate. However, compared to the standard Amber14SB simulation of 7K2G, the real-time unbound state sampling obtained with the advanced protocol was higher for simulations 2 and 3, meaning that the protocol is correcting the bias in the right direction. At the same time, the unbound state sampling is still lower than the one obtained with a99SB-disp (between 14% and 18%) and so is the resulting accuracy.

Given these results, the main direction to improve the preliminary protocol is to

find a way to increase the $6 \leftrightarrow 7$ exchange rate, for example by optimizing the choice of internal conformational CVs in order to maximize the potential energy overlap between the two replicas. Additional strategies to address the overlap problem, together with possible extensions of this approach, are discussed in chapter 4.

Chapter 4

Conclusions and Future Perspectives

This thesis addressed two distinct but related challenges in the prediction of absolute binding free energies (ΔG_b°) for protein-cyclic peptide complexes: overcoming the sampling limitations for binding-unbinding processes and the systematic bias introduced by different force fields on peptide exploration in the unbound state. To address the first challenge we built a multi-replica setup, combining OneOPES with funnel metadynamics, biasing two main collective variables (the funnel projection `pp.proj` and the native contact map `cmap`) while progressively adding hydration CVs along the replicas. This protocol was applied to six members of the 7K2* family and the structurally distinct 1SFI complex. In order to address the second challenge, we simulated the complexes both with Amber14SB and a99SB-disp, showing the introduction of the two opposite biases: the over-compaction of the free peptide in the unbound state caused by Amber14SB and the under-binding of rigid, pre-organized peptides caused by a99SB-disp. In particular we showed that the two errors are not symmetric in magnitude and that, while the Amber14SB error strongly correlates with the intrinsic flexibility of the free peptide (see figure 3.5), the a99SB-disp error does not show a statistically robust correlation (see figure 3.11). Based on the resulting dichotomy, we developed a three-level screening protocol which uses explicit and implicit solvent REMD simulations, combined with a KDE-based classification to assign each complex to the optimal force field (or force field combination, following the hybrid approach). We applied this protocol to the 7K2* family, obtaining a substantial improvement in the correlation between simulated and experimental ΔG_b° (Pearson $r = 0.85 \pm 0.09$) and in the MAE, which was around the 1 kcal/mol target, compared to results obtained with either isolated force fields in sections 3.1.1 and 3.1.2. Extending the protocol to 1SFI, we see that it generalizes reasonably well to a structurally unrelated, more rigid complex (outside of the 7K2* family), although we get a slightly reduced Pearson correlation.

Moreover, we established two complementary convergence criteria: the stabilization of $\Delta G_b^\circ(t)$, presented in section 2.4.1, across independent simulations below a 1 kcal/mol cross-replica standard deviation, and the stationarity of the free energy surface through the FES-RMSD evaluated against a 0.6 kcal/mol threshold and presented in section 2.4.2. The most evident conclusion of these convergence analyses is that at the simulation lengths currently accessible to this work (at most 500 ns for 1SFI), no system can be declared definitively converged, even when the ΔG_b° agrees well with experiments. Note, moreover, that the length of current blocks has been chosen as a compromise because of the limited simulation length and it produces an important simplification:

if successive blocks are not statistically independent, convergence analyses which compare successive blocks, like FES-RMSD, can underestimate the true uncertainty. That is why, once simulations are extended, the autocorrelation function of `pp.proj` will be computed for every system to find the correlation time and use it to make sure that the block length guarantees truly independent blocks, making the convergence analyses more rigorous.

Finally, we presented preliminary results of an advanced protocol which aims to overcome the force field choice altogether, by replacing the eighth replica of the standard setup with one containing only the free peptide in the unbound state, while biasing two internal conformational CVs. The advanced protocol was applied to 7K2G, simulated with Amber14SB, and showed a slight improvement in accuracy with respect to the standard multi-replica setup but its main limitation remains the low replica-exchange between the peptide-only and the complex replicas. Solving this issue remains the biggest challenge to improve this approach.

There are several extensions starting from the results obtained in this thesis, divided in short-term and long-term goals.

Short-term goals. As we have discussed, none of the systems we have studied has reached convergence according to the two criteria we have introduced in sections 2.4.1 and 2.4.2. Then, the most direct goal is to extend every simulation to at least 1 μs and apply again both convergence criteria, defining the block length according to the autocorrelation time of `pp.proj`, in order to see if the additional sampling time is sufficient to reach convergence. The second short-term goal is extending the current dataset by considering new complexes. As we can see in the correlation plots in chapter 3, we have two regions to fill: the first one is at high binding affinity, between that of 1SFI and the strongest binders of the 7K2* family and the other one at low affinities, around the weakest binders of the 7K2* family. Both of these gaps limit how much the correlation analysis can be trusted outside the affinity range actually sampled. Moreover, new complexes would be useful to extend the screening protocol and make it more robust.

Long-term goals. There are two long-term directions to extend this project: addressing the limitations and extending the advanced protocol (presented in section 3.5) and extending the three-level screening protocol (presented in section 3.2).

- The central limitation we found for the advanced protocol was the low replica exchange rate (0.6%-1.3%) between replicas 6 and 7. This is caused by the low overlap between the two replicas, which differ both in the biased CVs and in the physical system they contain (the whole complex for replica 6 and the isolated peptide for replica 7). A possible way to improve this is to change the architecture of the multi-replica setup, going from the current 7+1 architecture to a 4+4 architecture. Because the replica exchange rate observed for each system for the complex-only replicas (0-6) is consistently high (between 20% and 60%) we can dedicate fewer replicas to the whole complex and more to the isolated peptide. With a 4+4 architecture, the first block of replicas (0-3) will be dedicated to the whole complex and it will carry the same CVs of current replicas 0-6, while the second block will be entirely dedicated to the isolated peptide in the unbound state, with internal conformational CVs progressively added across replicas 4-7. In this way the temperature in replica 3 (the last replica with the whole

complex) and replica 4 (the first with the isolated peptide) will be lower, and the exploration of the unbound state should improve, since we can bias more internal conformational CVs of the peptide. An alternative to this approach, if exchange rate between replicas 0-3 decreases too much, would be to consider a 8+8 architecture, with 8 replicas dedicated to the entire complex (as in the original setup, see 2.2) and 8 replicas dedicated to the isolated peptide.

- The three-level screening protocol, presented in section 3.2, relies on the assumption that the Δ_{R_G} and Δ_{rmsd} distributions are bimodal, to find a data-driven cut-off between the two modes and assign each complex to a force field, or to the hybrid approach. As we have seen, however, the bimodality breaks down when more data points are added (see table 3.4 and 3.5). This suggests that an extension of the screening protocol is needed as the dataset grows and the distribution becomes more continuous. The idea is to replace the binary force field assignment with a continuous weight $w \in [0, 1]$, which should express how much a given complex should be described by a99SB-disp or Amber14SB based on the flexibility of the free peptide: $w = 0$ would correspond to a rigid peptide and would be assigned to Amber14SB, while $w = 1$ would correspond to a purely flexible one and would be assigned to a99SB-disp. Any intermediate value of w would simply be a weighted combination of the two force fields' ΔG_b° estimates. Another issue is that, right now, the screening protocol classifies flexibility using two distinct coefficients of variation, Δ_{R_G} for Level 2 and Δ_{rmsd} for Level 3. A promising long-term direction could be to replace both metrics with a single, physically more interpretable one: the conformational entropy of the joint rmsd-R_G distribution sampled by the isolated peptide. This entropy is composed of two complementary terms, $S = S_{\text{between}} + S_{\text{within}}$. The first term is:

$$S_{\text{between}} = -\sum_{k=1}^{K} p_k \ln p_k \tag{4.1}$$

which represents Shannon entropy where k is the current cluster, K is the total number of clusters identified in the rmsd-R_G plane, $p_k = N_k/N$ is the probability that a certain frame belongs to cluster k, with N_k the number of trajectory frames and N the total number of frames. This term captures the global flexibility of the peptide: a peptide populating many clusters is considered to be more flexible than one populating fewer clusters. However, S_{between} alone does not take into account how broad the cluster is, so two peptides with the same number of clusters and the same population per cluster would have the same entropy value, even if the clusters have different width. To correct for this effect, we add a second term:

$$S_{\text{within}} = \sum_k p_k S_k, \qquad S_k = \frac{1}{2} \ln \left[(2\pi e)^d |\Sigma_k| \right] \tag{4.2}$$

where S_k is the internal entropy of cluster k, d is the spatial dimension, e is the Euler's number and Σ_k is the covariance matrix, following the approach presented by Karplus and Kushick [75]. The resulting entropy, $S = S_{\text{between}} + S_{\text{within}}$ would replace both flexibility measures used in section 3.2.2 with a single descriptor. It is also plausibly easier to generalize into the continuous weight scheme, with $w \in [0, 1]$, discussed above. The application and validation of this entropy-based flexibility measure is left for future work.

Appendix A

Computational Techniques

A.1 PLUMED: The Core Engine

The collective variable definitions and enhanced sampling techniques (e.g., OneOPES, funnel metadynamics) used in this thesis are implemented through PLUMED [51], an open-source plug-in that provides some of the most popular MD codes with a common library of enhanced sampling algorithms and CVs, that we interface directly with GROMACS [50].

Theoretical background. MD simulations produce the trajectories of a large number of atoms, meaning the final result is a high-dimensional description of how atoms positions change as a function of time. It can be very difficult to interpret these results and compare them with experimental data without further processing the trajectories. Then, a convenient way of processing these data is to project the trajectory onto a small number of collective variables $\mathbf{s}(\mathbf{X}) = \{s_1(\mathbf{X}), \dots, s_n(\mathbf{X})\}$ and compute the associated probability distribution $P(\mathbf{s})$, from which the free energy follows: $F(\mathbf{s}) = -k_B T \log P(\mathbf{s})$. The equation implicitly assumes that the trajectory is ergodic, thus there are no high-energy barriers that prevent the system from visiting all the accessible configuration space. If this is not the case, PLUMED allows to add a bias potential $V(\mathbf{s}(\mathbf{X}), t)$ to enhance sampling. Since the bias depends on the atomic coordinates only through the CVs, then the corresponding force on each atom can be obtained through the chain rule:

$$\frac{\partial V}{\partial x_m} = \sum_{i=1}^{n} \frac{\partial V}{\partial s_i} \frac{\partial s_i}{\partial x_m} \tag{A.1}$$

Following equation A.1, the PLUMED code can be divided into units that calculate the CVs and their derivatives, functions of CVs and biases [51].

Architecture and interface with GROMACS. PLUMED is organized around objects called "Actions": every single line of a PLUMED input file can be cast as an Action object, whether it computes a CV (DISTANCE, ANGLE, TORSION), combines quantities to create a new CV, applies a bias potential or simply prints output. Actions cannot be performed in an arbitrary order as Actions that take, for example, CVs as an input cannot be computed without first calculating the CVs. Then, PLUMED automatically orders Actions according to their mutual dependencies [51]. This characteristic makes the construction of complex CVs (such as the funnel projection or

hydration CVs used in this thesis) particularly straightforward. The communication between PLUMED and GROMACS is also rather straightforward as PLUMED is called from within GROMACS at three points: once during initialization, once after each force evaluation and once at the end of the simulation. Moreover, since PLUMED is compiled independently from GROMACS, it can be patched into GROMACS (as we will use it), or it can be used as a standalone post-processing tool. Finally, PLUMED natively supports On-the-fly Probability Enhanced Sampling (OPES) [54], which combined with Hamiltonian Replica-Exchange (HREX) [49] constitute the foundation of OneOPES.

A.2 Hamiltonian Replica Exchange (HREX) Architecture

Standard replica-exchange methods, like parallel tempering, simulate several replicas of the same system at different temperatures, and coordinate exchanges are regulated through a Monte Carlo criterion. However, since the entire system needs to be accelerated [63], then the number of replicas needed to span a required range of temperatures increases with the size of the system, making these methods rather costly for protein-cyclic peptide complexes. Hamiltonian Replica Exchange (HREX) methods, such as REST2 [64], allow different replicas to evolve according to different Hamiltonians, and they are typically more efficient than parallel tempering, as the number of required replicas is much less [49]. In this thesis we use the HREX implementation for GROMACS introduced by Bussi [49], which allows different Hamiltonians to be assigned to different replicas and, more importantly, it is designed to work with PLUMED, so that we can combine HREX and OPES. In this implementation at each attempted exchange step, the coordinates of two neighbouring replicas are unconditionally exchanged and the total potential energy, which is the sum of the force field potential and of additional bias computed in PLUMED, is computed under both Hamiltonians. Then, this energy is stored for later usage and original coordinates are restored through another unconditional exchange. Finally, at the end of the MD step, when the actual exchange is attempted, the previously stored energy is used to accept or reject the exchange, using the same Metropolis-Hasting criterion introduced for OneOPES (2.9).

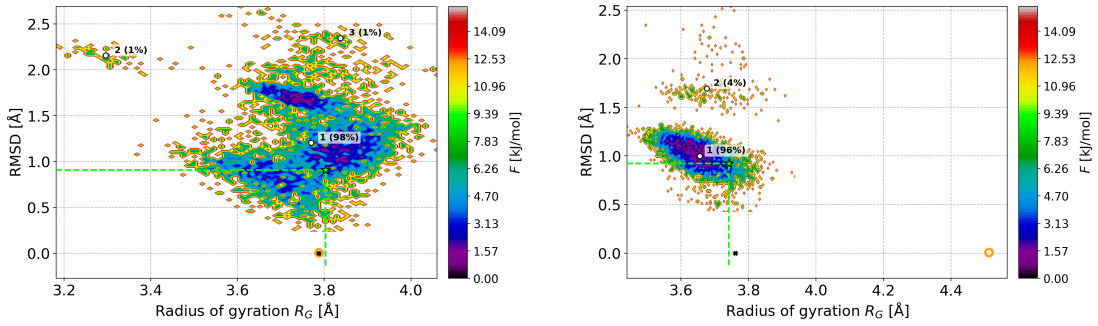
Relevance to ergodicity. It is important to note that, using the plain temperature replica-exchange scheme natively available in GROMACS, would not correctly handle exchanges between different replicas as the acceptance criterion is only valid when the Hamiltonian is common for all replicas, or if differs by temperature alone. In fact, using plain temperature replica-exchange, there is no communication between GROMACS and PLUMED, meaning that when an exchange is attempted the energy difference does not take into account the bias potentials computed by PLUMED, destroying the ergodicity of the exchange procedure. This is ultimately why HREX needs to be used within the simulation setup.

Appendix B

Conformational sampling

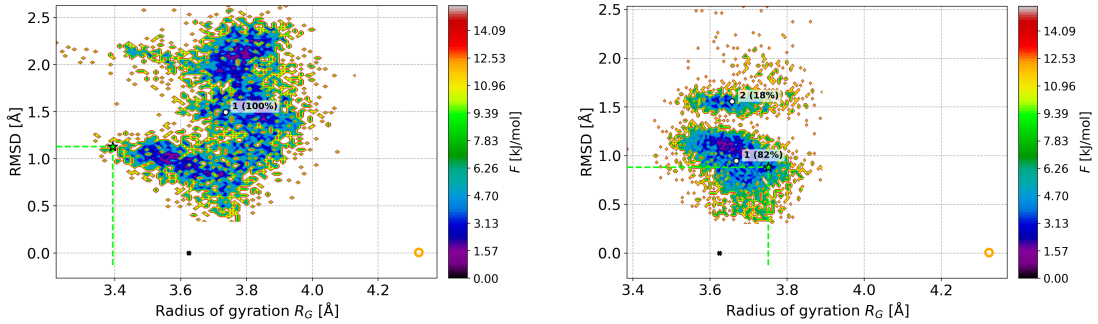
In this appendix we are going to report the joint distributions of rmsd and radius of gyration both in explicit and implicit solvent for all systems studied in this thesis, as we have already done for complex 7K2G in section 3.2.1. As we have seen for 7K2G, also for explicit solvent REMD simulations of the remaining complexes we have a single basin, continuous cluster. There are, however, different behaviours in the implicit solvent REMD simulations, depending on the peptide flexibility:

- 7K2F and 7K2E, as 7K2G, have a multiple basin map, while in the explicit solvent simulation they present a single basin, continuous cluster, as a consequence of over-compaction.
- Peptides classified as relatively ordered (the ones assigned to the hybrid approach by the three-level screening protocol) present a single basin also in the implicit solvent simulation, meaning explicit and implicit solvent sample more or less the same conformational space for relatively ordered peptides. The only exception is 7K2E, which has been assigned to the hybrid approach but shows two clusters in figure B.1b. This is not completely surprising as 7K2E is the complex, in the 7K2* family, with the highest Δ_{rmsd} value in Level 3 of the screening protocol (see section 3.2.2).
- 1SFI (figure B.6) has a different behaviour with respect to the 7K2* family: it has a single basin map in the explicit solvent simulation, as expected from a rigid complex, but it shows two separate clusters in the implicit solvent one: the rightmost cluster is the one present also in the explicit solvent simulation, but there is one more at low values of R_G and high values of rmsd. This is an interesting behaviour and could be caused by temperature, since implicit solvent simulations go up to 700 K. To check this argument, we did a 50 ns implicit solvent REMD simulation of 1SFI with $T_{\text{max}} = 455K$ and found out that the conformational space sampled by the peptide reduces again to the single cluster observed in the explicit solvent simulation, as can be seen in figure B.7.



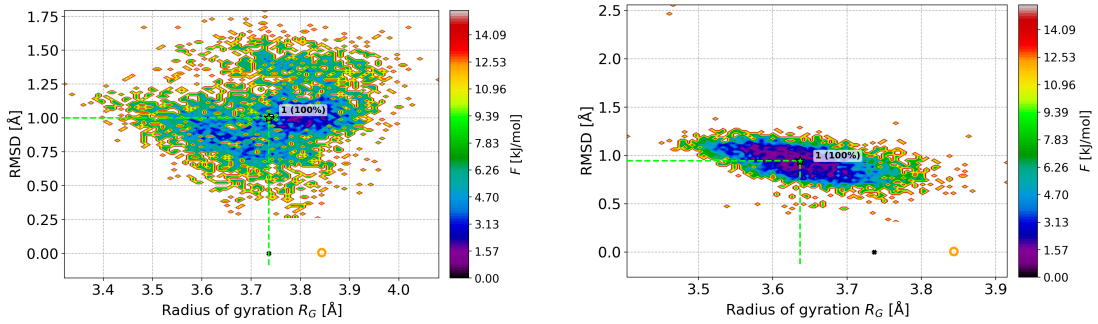
(a) explicit (b) implicit

Figure B.1: Two-dimensional free energy surface projected onto the backbone rmsd and radius of gyration sampled by 7K2E in the 50 ns REMD simulations both in explicit and implicit solvent.



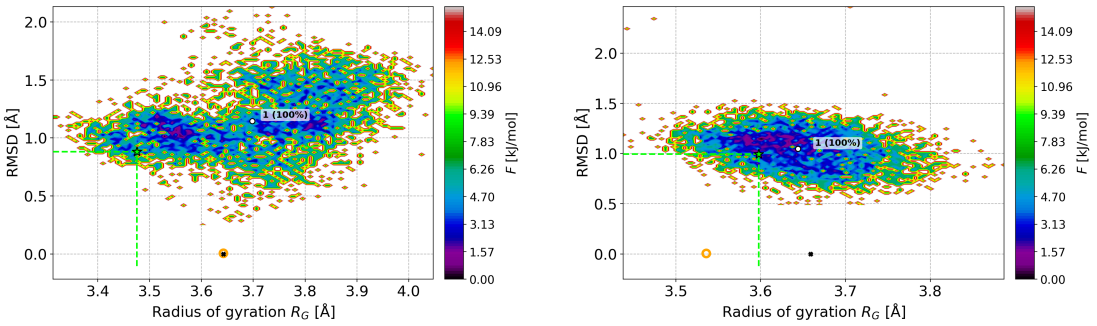
(a) explicit (b) implicit

Figure B.2: Two-dimensional free energy surface projected onto the backbone rmsd and radius of gyration sampled by 7K2F in the 50 ns REMD simulations both in explicit and implicit solvent.



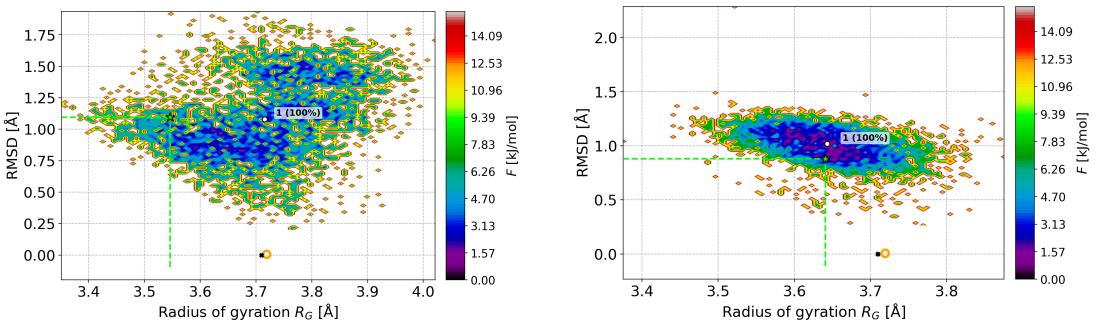
(a) explicit (b) implicit

Figure B.3: Two-dimensional free energy surface projected onto the backbone rmsd and radius of gyration sampled by 7K2H in the 50 ns REMD simulations both in explicit and implicit solvent.



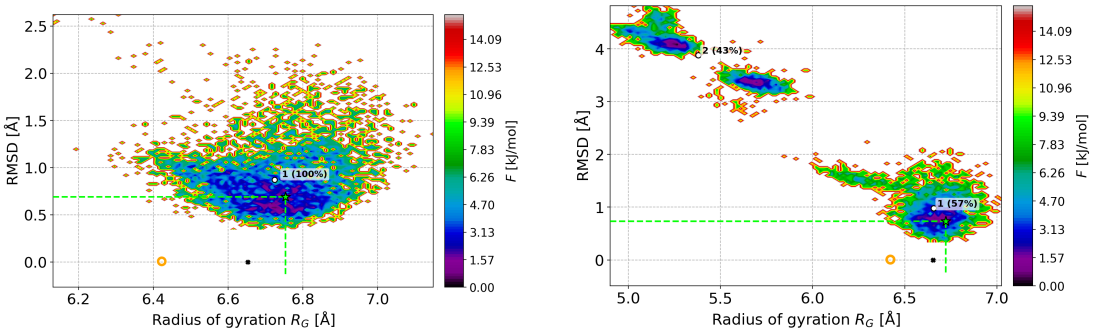
(a) explicit (b) implicit

Figure B.4: Two-dimensional free energy surface projected onto the backbone rmsd and radius of gyration sampled by 7K2I in the 50 ns REMD simulations both in explicit and implicit solvent.



(a) explicit (b) implicit

Figure B.5: Two-dimensional free energy surface projected onto the backbone rmsd and radius of gyration sampled by 7K2M in the 50 ns REMD simulations both in explicit and implicit solvent.



(a) explicit (b) implicit

Figure B.6: Two-dimensional free energy surface projected onto the backbone rmsd and radius of gyration sampled by 1SFI in the 50 ns REMD simulations both in explicit and implicit solvent.

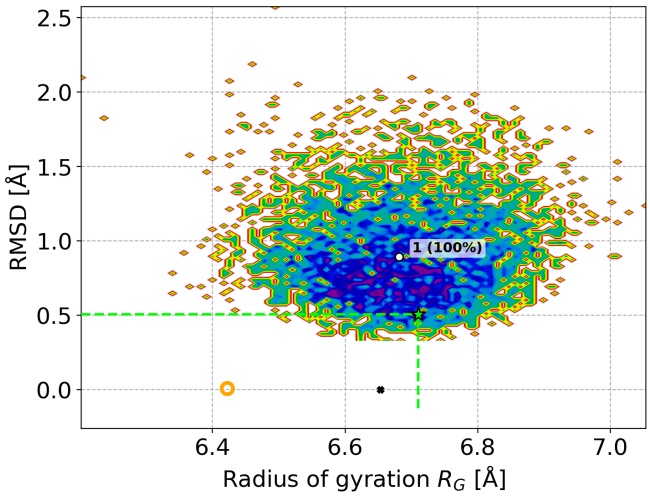


Figure B.7: Two-dimensional free energy surface projected onto the backbone rmsd and radius of gyration sampled by 1SFI in the 50 ns REMD simulation in implicit solvent with $T_{\max} = 455K$.

Appendix C

Kernel Density Estimation

Kernel density estimation (KDE) [74] is a well-known nonparametric approach used to estimate the underlying probability density function of a dataset, requiring no assumptions that the underlying density function is from a parametric family [76]. Then, KDE learns the shape of the density function directly from the data. Let us consider $X_1, \ldots, X_n \in \mathbb{R}^d$ an i.i.d. random sample, the KDE can be expressed as follows:

$$\hat{p}_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{C.1}$$

where K is a smooth function called the Kernel function and $h > 0$ is the smoothing bandwidth. Two common examples of Kernels are the Gaussian Kernel (used in this thesis) and the spherical Kernel. Looking at eq. C.1, we can see KDE has the effect of treating each data point as a smooth bump, whose shape is determined by the Kernel function $K(x)$. Then, KDE sums over all these bumps to get a density estimator. An example with six data points can be seen in figure C.1 below. As presented in section

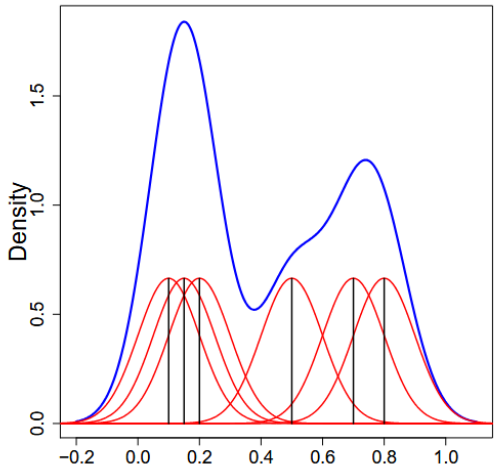


Figure C.1: 1d illustration of how KDE is constructed. Each observation is located where the black lines are, we smooth those into small (red) bumps and then sum them to obtain a density estimator (blue curve). Adapted from Chen (2017) [76].

3.2, our screening protocol employs a one-dimensional Gaussian Kernel and while the specific choice of it does not play an important role in KDE, the choice of the bandwidth

h is a critical one [76]. It controls the amount of smoothing applied to the data: if h is too small then the density estimate could have numerous artificial peaks, giving a noise estimate, if h is too large the estimator will oversmooth the data, possibly erasing important features of the distribution. To balance this trade-off and maintain the separation between physically different systems we empirically fixed the bandwidth to $h = 0.5\ \sigma_{\text{data}}$.

KDE can also be used to estimate geometric (and topological) structures related to density estimate. In particular it is used to identify local modes, which are defined as points where the density is locally maximized and its gradient is zero. In the context of this thesis, local modes correspond to the highly populated clusters (for example the disordered and ordered groups in figure 3.15) along the Δ_{rmsd} dimension. By mathematically identifying the local minimum between these two local modes, the screening protocol defines a cut-off value ($\Delta_{\text{rmsd}}^{\text{cut-off}}$) that physically separates different complexes and assigns them to their optimal force field (or force field combination).

Appendix D

Leave-One-Out Cross Validation of the Screening Protocol

The predictive screening protocol we have introduced in section 3.2, uses KDE to define a separation cut-off based on the intrinsic flexibility of the cyclic peptides measured through Δ_{rmsd}. Because of the low number of data present in this analysis, it is important to evaluate the statistical robustness of the derived cut-offs against the addition or removal of data points.

Here, we consider a Leave-One-Out cross validation for Level 3 (implicit solvent REMD) of the protocol by removing one peptide from the dataset and recomputing the KDE and its local minimum, to show that the derived cut-off remains stable. The results obtained by removing one of the peptides from the analysis are summarised in the table D.1 below. As we can see, removing any member of the 7K2* family

Omitted System	**Implicit Solvent Cut-off**
None (All systems)	0.267
7K2E	0.255
7K2F	0.285
7K2G	0.263
7K2H	0.265
7K2I	0.261
7K2M	0.265

Table D.1: Leave-One-Out cross-validation of the KDE cut-offs. The table reports the estimated cut-off values for Level 3 analysis when a single peptide from the 7K2* family is iteratively removed from the dataset.

doesn't change greatly the cut-off as it remains stable around the value of 0.267. Most importantly, the classification ordered vs disordered of the remaining peptides is not altered during the analysis meaning that, even if we had removed one peptide from the dataset, the complexes would still be assigned to the same force fields (or to the hybrid approach).

Appendix E

Screening Protocol with Alternative Flexibility Metrics

As we have seen in section 3.2.2, the choice of Δ_{R_G} for Level 2 (across different families), and of Δ_{rmsd} for Level 3 (within the same family) is not arbitrary and it is justified during the thesis. Here we want to show what happens when we swap the two descriptors, confirming our theoretical expectations.

Δ_{rmsd} at Level 2. We report in Table E.1 the mean and standard deviation of the backbone rmsd, together with the coefficient of variation Δ_{rmsd} of all seven complexes, obtained from the same 50 ns explicit-solvent REMD simulation we have used for the main text Level 2 analysis.

System	μ_{rmsd} (Å)	σ_{rmsd} (Å)	Δ_{rmsd}
7K2E	1.2426	0.4382	0.3526
7K2F	1.4905	0.5212	0.3497
7K2G	1.4117	0.3904	0.2765
7K2H	1.0089	0.2468	0.2447
7K2I	1.1502	0.2742	0.2384
7K2M	1.0827	0.3130	0.2891
1SFI	0.9179	0.3434	0.3741

Table E.1: Mean, standard deviation and coefficient of variation of the backbone rmsd from 50 ns explicit-solvent REMD simulations, for all seven complexes.

Figure E.1 shows the resulting KDE distribution, from which we can see that, while the distribution is still formally bimodal, 1SFI, which should be well separated from the 7K2* family, is now grouped together with 7K2E and 7K2F.

Δ_{R_G} at Level 3. In table E.2 we report the same information reported in the previous paragraph (but now only for the 7K2* complexes), obtained from the same 50 ns implicit-solvent REMD simulations of the isolated peptides used for the main text analysis of Level 3.

Figure E.2 shows the resulting KDE distribution. We can see that only 7K2G is above the cut-off, contradicting the theoretical expectation and the result obtained in section 3.2.2 at Level 3. In fact, 7K2F, which exactly like 7K2G lacks a stabilizing side chain, is grouped together with three ordered, E78P-substituted peptides (7K2I, 7K2H,

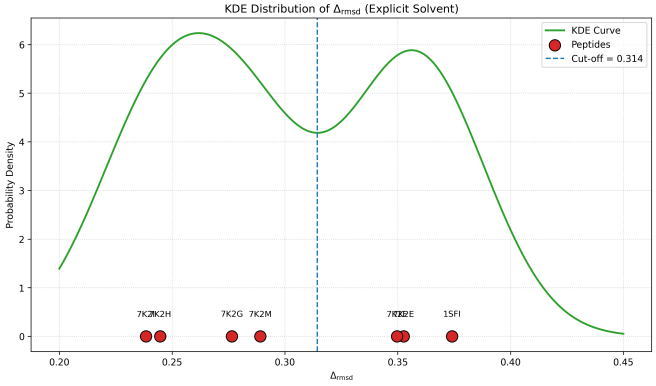


Figure E.1: KDE distribution of Δ_{rmsd} obtained from the same 50 ns explicit-solvent REMD simulations used for Level 2. The grouping does not correctly separate 1SFI from the 7K2* family.

System	μ_{R_G} **(Å)**	σ_{R_G} **(Å)**	Δ_{R_G}
7K2E	3.6563	0.0677	0.0185
7K2F	3.6681	0.0702	0.0191
7K2G	3.6222	0.0932	0.0257
7K2H	3.6400	0.0638	0.0175
7K2I	3.6449	0.0663	0.0182
7K2M	3.6435	0.0639	0.0175

Table E.2: Mean, standard deviation and coefficient of variation of the radius of gyration from 50 ns implicit-solvent REMD simulations, for the six 7K2* peptides.

7K2M). This result confirms our theoretical expectation [73] that R_G, since it is not sensitive to point substitutions, cannot correctly distinguish the flexibility of peptides that have identical size and topology.

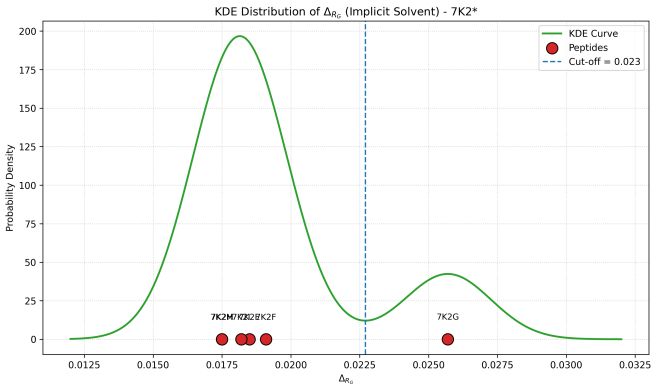


Figure E.2: KDE distribution of Δ_{R_G} obtained from the same 50 ns implicit-solvent REMD simulations used for Level 3 in the main text. The resulting grouping does not match the expected ordered/disordered classification within the 7K2* family.

Appendix F

Computational details

In this section we summarize the simulation pipeline, which is common for all protein-cyclic peptide complexes studied in this thesis. Simulation lengths and equilibration times can vary depending on the system and they are reported in the respective sections along chapter 3 but the overall protocol, from system preparation to production runs, is the same for the 7K2* complexes and 1SFI.

System preparation and MD parameters. We took the crystal structure of each complex from the benchmark dataset of Zhao *et al.* [22] and the topology was generated using GROMACS, selecting the appropriate force field (Amber14SB or a99SB-disp) with its associated water model (TIP3P or a modified TIP4P-D-like model, also called a99SB-disp water). Then the complex was put in a dodecahedral box, solvated, neutralized and energy minimization was performed using a steepest-descent algorithm. MD simulations were integrated using the leap-frog algorithm with a 2 fs time step, constraining hydrogen bonds through the LINCS algorithm. Temperature was kept at $T = 300K$ through the stochastic velocity-rescaling thermostat (`v-rescale`) using a time constant of 0.1 ps. Pressure was instead kept at 1 bar using the `C-rescale` barostat, with a time constant of 1 ps and isotropic scaling. An important difference between force fields was in the non-bonded interaction parameters: simulations with Amber14SB used a 1.0 nm cut-off for both Coulomb and Van Der Waals interactions, coupled with standard analytical dispersion corrections for energy and pressure (`DispCorr = EnerPres`). On the other hand, simulations with a99SB-disp used a 1.2 nm cut-off for non-bonded interactions, while disabling analytical dispersion corrections (`DispCorr = no`), following the original force field parametrization [42, 43]. Particle Mesh Ewald (PME) was used in all simulations to take care of long-range electrostatics.

Equilibration. After energy minimization, every system was equilibrated in two steps: a 5 ns run in the NVT ensemble, followed by a 10 ns run in the NPT ensemble. Position restraints were applied to the heavy atoms of the protein-cyclic peptide complex to allow for proper solvent relaxation. The production runs were then carried out in the NPT ensemble without any position restraints.

CVs construction. Once equilibration was completed, the CVs described in 2.3 were built. In particular a 100 ns unbiased simulation of the whole complex was performed, from which the standard deviation of each CV was estimated from the resulting trajectory using PLUMED. These values were used to set the initial Kernel width (`SIGMA`)

of the OPES bias, so that the initial Gaussians followed the natural fluctuation scale of the complex. Finally, the CVs definitions were encoded in different PLUMED input files, one for each replica, with the number of biased CVs increasing from replica 0 to replica 7, as described in table 2.2.

OneOPES parameters. There are three parameters that drive the behaviour of each OPES Explore bias: `BARRIER`, `SIGMA` and `PACE`. `BARRIER` should provide an upper estimate of the free energy barrier the bias must overcome. This parameter was the same for all 7K2* complexes, despite their different binding free energies (see 2.1), as using a sufficiently large `BARRIER` parameter avoids a system-dependent choice, although a different `BARRIER` parameter was used for auxiliary hydration CVs (10 times smaller than the one set for main CVs), as we expect that the free energy barrier across these transversal degrees of freedom is considerably lower than the one along main CVs. `SIGMA` is the initial Kernel width and, as said before, was set based on unbiased fluctuations. The `PACE`, which is the frequency at which new kernels are deposited, was kept fixed for every system and it was doubled for auxiliary CVs with respect to main CVs. The `PACE` was chosen as a compromise: if the bias is deposited too frequently, the peptide could go out of the binding pocket before it has been thoroughly explored. On the other hand, if the bias deposition is too slow, we could have really slow convergence.

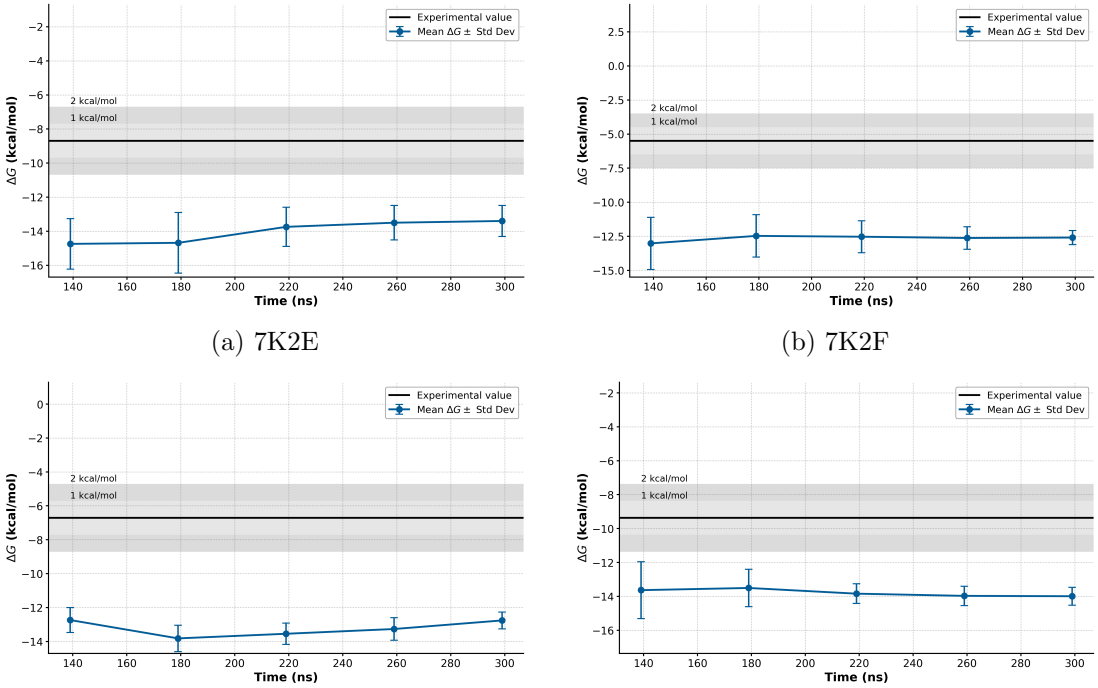
After building the PLUMED files for each replica, containing the CVs, funnel parameters and OneOPES parameters, the eight replicas were simulated in parallel using GROMACS 2023.5 patched with PLUMED 2.9.4, with Hamiltonian Replica Exchange coordinate swaps attempted every 2000 steps.

Appendix G

Simulations Results

G.1 Results with Amber14SB

In this section we present, schematically, the same set of analyses presented in section 3.1.1 for 7K2H and 7K2G, for the remaining systems (7K2E, 7K2F, 7K2I, 7K2M): the five-point $\Delta G_b^\circ(t)$ analysis, the FES-RMSD between successive 40 ns blocks, the cumulative and block averaged real-time sampling of the unbound state and the mean 1d FES along the funnel projection axis.



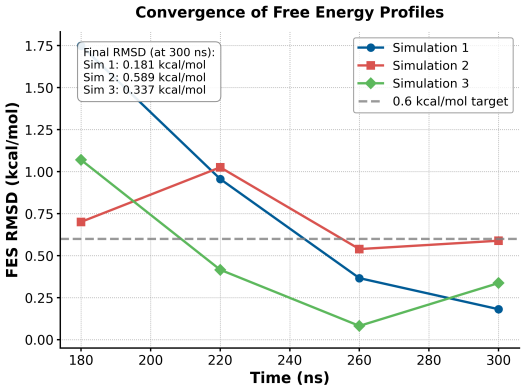
(a) 7K2E

(b) 7K2F

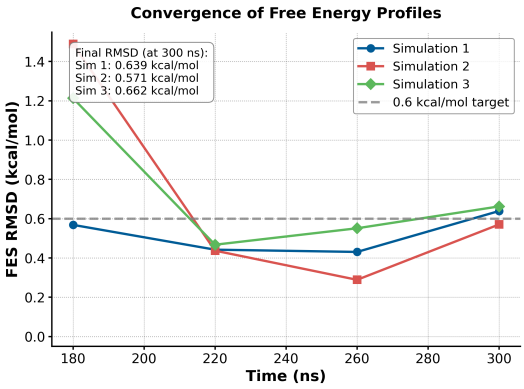
(c) 7K2I

(d) 7K2M

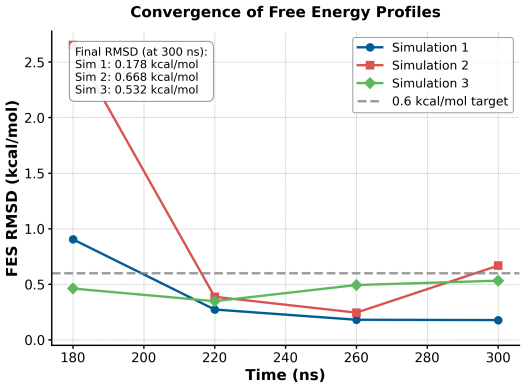
Figure G.1: Absolute binding free energy $\Delta G_b^\circ(t)$ (mean ± std over three independent simulations), evaluated at five sparse time points for 7K2E, 7K2F, 7K2I and 7K2M with Amber14SB. Black lines mark the experimental values.



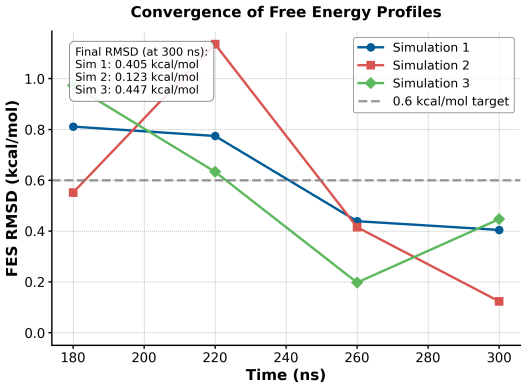
(a) 7K2E



(b) 7K2F

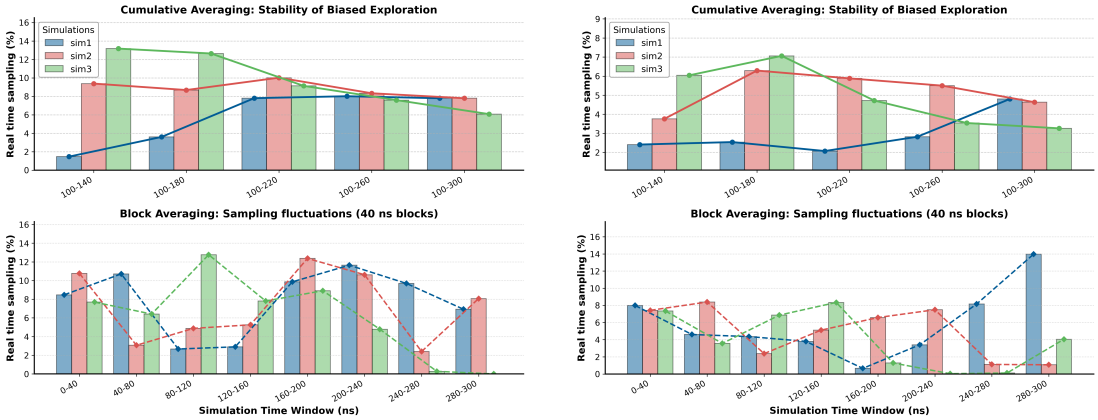


(c) 7K2I



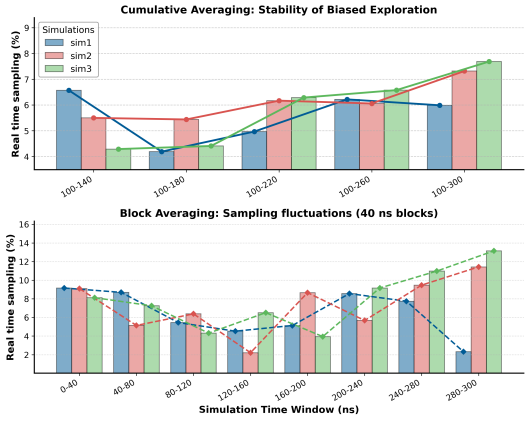
(d) 7K2M

Figure G.2: FES-RMSD between successive 40 ns blocks for three independent simulations of 7K2E, 7K2F, 7K2I and 7K2M with Amber14SB. The dashed lines represent the 0.6 kcal/mol target.

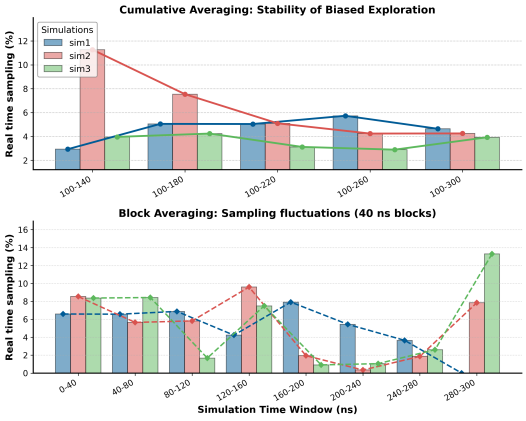


(a) 7K2E

(b) 7K2F

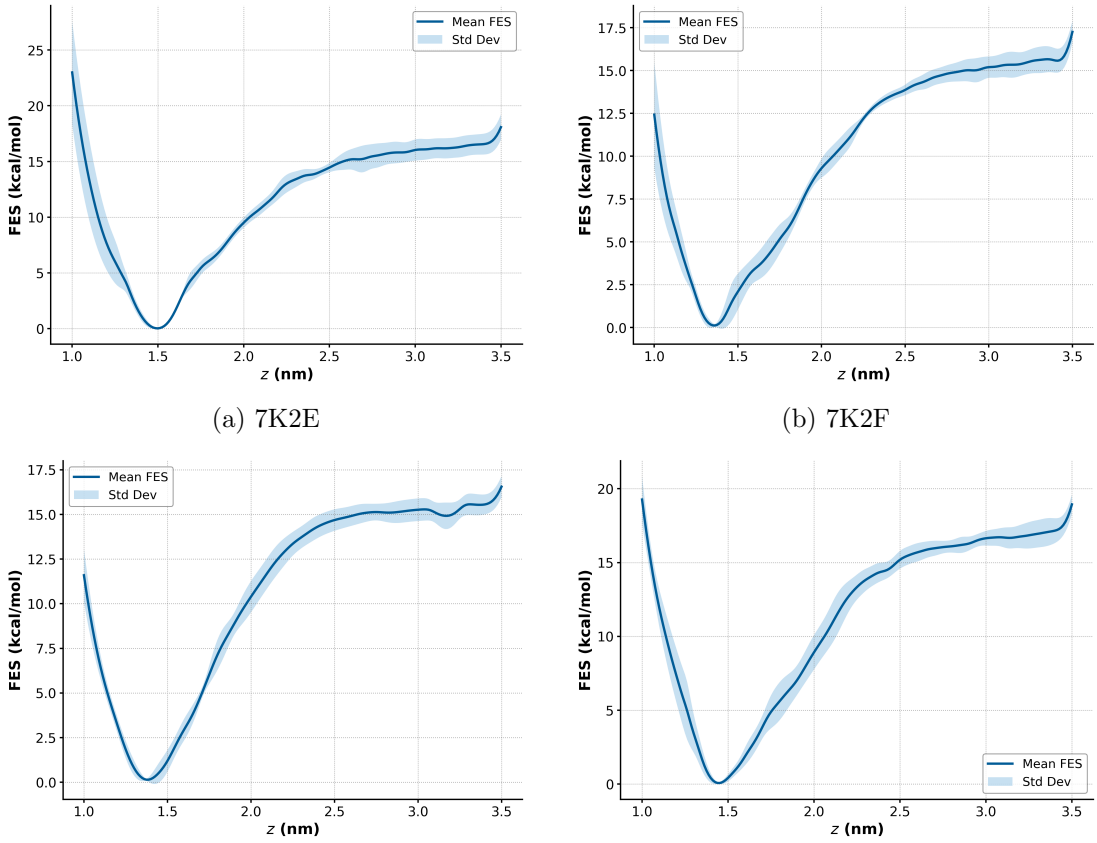


(c) 7K2I



(d) 7K2M

Figure G.3: Cumulative (top) and block-averaged (bottom) real-time sampling of the unbound state for three independent simulations of 7K2E, 7K2F, 7K2I and 7K2M with Amber14SB.



(a) 7K2E

(b) 7K2F

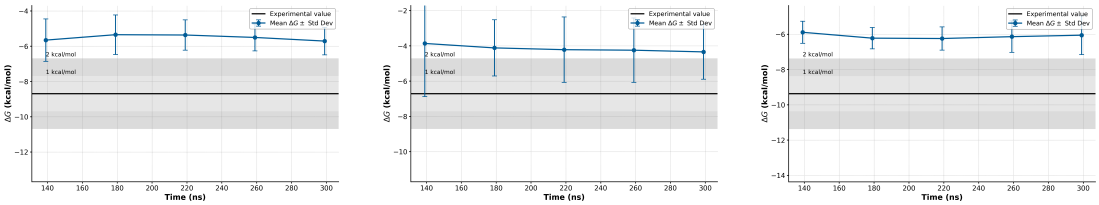
(c) 7K2I

(d) 7K2M

Figure G.4: 1d Free Energy Surfaces (FES) along the funnel projection axis for 7K2E, 7K2F, 7K2I and 7K2M with Amber14SB.

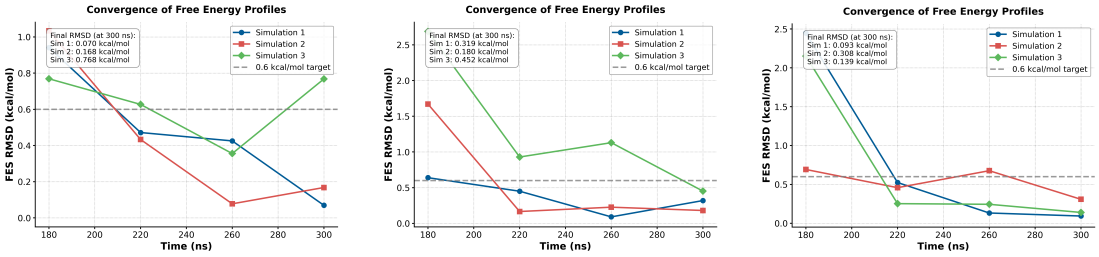
G.2 Results with a99SB-disp

In this section we report the analyses presented in section 3.1.2 with a99SB-disp for 7K2H and 7K2G. Note that we present here the results for 7K2E, 7K2I and 7K2M since the ones for 7K2F have been presented in section 3.3.



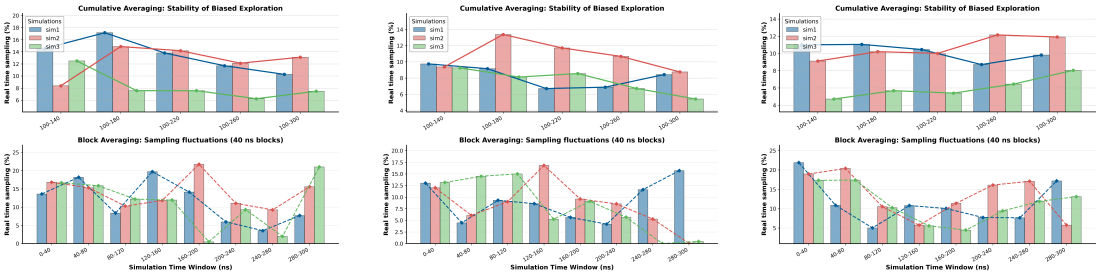
(a) 7K2E (b) 7K2I (c) 7K2M

Figure G.5: Absolute binding free energy $\Delta G_b^\circ(t)$ (mean $\pm$ std over three independent simulations), evaluated at five sparse time points for 7K2E, 7K2I, 7K2M with a99SB-disp. Black lines mark the experimental values.



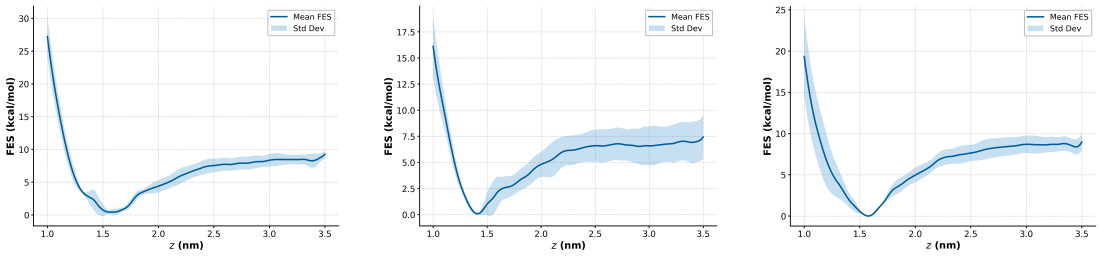
(a) 7K2E (b) 7K2I (c) 7K2M

Figure G.6: FES-RMSD between successive 40 ns blocks for three independent simulations of 7K2E, 7K2I and 7K2M with a99SB-disp. The dashed lines represent the 0.6 kcal/mol target.



(a) 7K2E (b) 7K2I (c) 7K2M

Figure G.7: Cumulative (top) and block averaged (bottom) real-time sampling of the unbound state for three independent simulations of 7K2E, 7K2I and 7K2M with a99SB-disp.

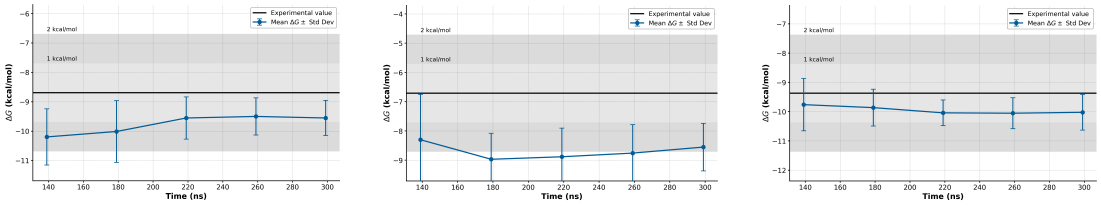


(a) 7K2E (b) 7K2I (c) 7K2M

Figure G.8: 1d Free Energy Surfaces (FES) along the funnel projection axis for 7K2E, 7K2I and 7K2M with a99SB-disp.

G.3 Application of the Screening Protocol

In this section we present the five-point $\Delta G_b^\circ(t)$ analysis for the remaining three systems assigned to the hybrid approach by the three-level screening protocol (7K2E, 7K2I, 7K2M) which complete the results already shown for 7K2H in section 3.3. This is the only convergence check available for these systems, since the FES-RMSD and population analyses cannot be defined, as explained in sections 3.2 and 3.3.



(a) 7K2E (b) 7K2I (c) 7K2M

Figure G.9: Absolute binding free energy $\Delta G_b^\circ(t)$ evaluated at five sparse time points for 7K2E, 7K2I, 7K2M obtained through the hybrid approach. Error bars are obtained by propagating the cross-replica standard deviations of the independent Amber14SB and a99SB-disp estimates. Black lines mark the experimental values.

Bibliography

[1] Hossam Nada et al. "New insights into protein–protein interaction modulators in drug discovery and therapeutic advance". In: *Signal Transduction and Targeted Therapy* 9.1 (Dec. 6, 2024), p. 341. DOI: 10.1038/s41392-024-02036-3.

[2] Emil Spreitzer, Sinem Usluer, and Tobias Madl. "Probing Surfaces in Dynamic Protein Interactions". In: *Journal of Molecular Biology* 432.9 (Apr. 2020), pp. 2949–2972. DOI: 10.1016/j.jmb.2020.02.032.

[3] Renata Varnaitė and Stuart A. MacNeill. "Meet the neighbors: Mapping local protein interactomes by proximity-dependent labeling with BioID". In: *PROTEOMICS* 16.19 (2016), pp. 2503–2518. DOI: 10.1002/pmic.201600123.

[4] Meng Gao, Kui Cheng, and Hang Yin. "Targeting protein-protein interfaces using macrocyclic peptides". In: *Peptide Science* 104.4 (2015), pp. 310–316. DOI: 10.1002/bip.22625.

[5] Samuel Murail et al. "Robust Conformational Space Exploration of Cyclic Peptides by Combining Different MD Protocols and Force Fields". In: *Journal of Chemical Theory and Computation* 21.19 (Oct. 14, 2025), pp. 10018–10034. DOI: 10.1021/acs.jctc.5c01123.

[6] John A. Robinson et al. "The design, structures and therapeutic potential of protein epitope mimetics". In: *Drug Discovery Today* 13.21 (Nov. 2008), pp. 944–951. DOI: 10.1016/j.drudis.2008.07.008.

[7] Elif Ozkirimli and Berna Sariyar. "Protein-Peptide Interactions Revolutionize Drug Development". In: *Binding Protein* (2012). DOI: 10.5772/48418.

[8] Lei Wang et al. "Therapeutic peptides: current applications and future directions". In: *Signal Transduction and Targeted Therapy* 7.1 (Feb. 14, 2022), p. 48. DOI: 10.1038/s41392-022-00904-4.

[9] Xinjian Ji, Alexander L. Nielsen, and Christian Heinis. "Cyclic Peptides for Drug Development". In: *Angewandte Chemie International Edition* 63.3 (2024), e202308251. DOI: 10.1002/anie.202308251.

[10] Michelle W. Y. Southey and Michael Brunavs. "Introduction to small molecule drug discovery and preclinical development". In: *Frontiers in Drug Discovery* 3 (Nov. 30, 2023). DOI: 10.3389/fddsv.2023.1314077.

[11] Aw Roberts and Dcs Huang. "Targeting BCL2 With BH3 Mimetics: Basic Science and Clinical Application of Venetoclax in Chronic Lymphocytic Leukemia and Related B Cell Malignancies". In: *Clinical Pharmacology & Therapeutics* 101.1 (2017), pp. 89–98. DOI: 10.1002/cpt.553.

[12] Patrick G. Dougherty, Ashweta Sahni, and Dehua Pei. "Understanding Cell Penetration of Cyclic Peptides". In: *Chemical Reviews* (2019). DOI: 10.1021/acs.chemrev.9b00008.

[13] Katja Zerbe, Kerstin Moehle, and John A. Robinson. "Protein Epitope Mimetics: From New Antibiotics to Supramolecular Synthetic Vaccines". In: *Accounts of Chemical Research* (2017). DOI: 10.1021/acs.accounts.7b00129.

[14] Timothy A. Hill et al. "Constraining Cyclic Peptides To Mimic Protein Structure Motifs". In: *Angewandte Chemie International Edition* 53.48 (2014), pp. 13020–13041. DOI: 10.1002/anie.201401058.

[15] Bhavesh Khatri, Venkateswara Rao Nuthakki, and Jayanta Chatterjee. "Strategies to Enhance Metabolic Stabilities". In: *Methods in Molecular Biology* (2019). DOI: 10.1007/978-1-4939-9504-2_2.

[16] Benjamin Leader, Quentin J. Baca, and David E. Golan. "Protein therapeutics: a summary and pharmacological classification". In: *Nature Reviews Drug Discovery* (2008). DOI: 10.1038/nrd2399.

[17] Paul McGonigle. "How Biologics Have Changed the Drug Discovery Landscape". In: *Annual Review of Pharmacology and Toxicology* 65.1 (Jan. 23, 2025), pp. 29–46. DOI: 10.1146/annurev-pharmtox-061724-080811.

[18] Braxten D. Hornsby et al. "Therapeutic peptides and proteins: Status and developments in drug delivery". In: *Journal of Controlled Release* 394 (June 2026), p. 114895. DOI: 10.1016/j.jconrel.2026.114895.

[19] Maria-Jesus Blanco and Kevin M. Gardinier. "New Chemical Modalities and Strategic Thinking in Early Drug Discovery". In: *ACS Medicinal Chemistry Letters* 11.3 (Mar. 12, 2020), pp. 228–231. DOI: 10.1021/acsmedchemlett.9b00582.

[20] Karen M. Corbett et al. "Cyclosporin Structure and Permeability: From A to Z and Beyond". In: *Journal of Medicinal Chemistry* (2021). DOI: 10.1021/acs.jmedchem.1c00580.

[21] Alessia Ghidini, Eleonora Serra, and Andrea Cavalli. "On Free Energy Calculations in Drug Discovery". In: *Accounts of Chemical Research* 58.20 (Oct. 21, 2025), pp. 3137–3145. DOI: 10.1021/acs.accounts.5c00465.

[22] Huifeng Zhao et al. "Improving the predictive performance of binding affinities and poses for protein–cyclic peptide complexes through fine-tuned MM/PBSA(GBSA)-based methods". In: *Briefings in Bioinformatics* 26.6 (Nov. 1, 2025), bbaf632. DOI: 10.1093/bib/bbaf632.

[23] Samuel Genheden and Ulf Ryde. "The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities". In: *Expert Opinion on Drug Discovery* 10.5 (May 4, 2015), pp. 449–461. DOI: 10.1517/17460441.2015.1032936.

[24] Gaoqi Weng et al. "Assessing the performance of MM/PBSA and MM/GBSA methods. 9. Prediction reliability of binding affinities and binding poses for protein–peptide complexes". In: *Physical Chemistry Chemical Physics* 21.19 (May 15, 2019), pp. 10135–10145. DOI: 10.1039/c9cp01674k.

[25] Benoît Roux and Christophe Chipot. "Editorial Guidelines for Computational Studies of Ligand Binding Using MM/PBSA and MM/GBSA Approximations Wisely". In: *The Journal of Physical Chemistry B* 128.49 (Dec. 2, 2024), pp. 12027–12029. DOI: 10.1021/acs.jpcb.4c06614.

[26] Ercheng Wang et al. "End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design". In: *Chemical Reviews* (2019). DOI: 10.1021/acs.chemrev.9b00055.

[27] Tiziano Tuccinardi. "What is the current value of MM/PBSA and MM/GBSA methods in drug discovery?" In: *Expert Opinion on Drug Discovery* 16.11 (Nov. 2, 2021), pp. 1233–1237. DOI: 10.1080/17460441.2021.1942836.

[28] G. M. Torrie and J. P. Valleau. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling". In: *Journal of Computational Physics* (1977). DOI: 10.1016/0021-9991(77)90121-8.

[29] Alessandro Laio and Michele Parrinello. "Escaping free-energy minima". In: *Proceedings of the National Academy of Sciences* 99.20 (Oct. 2002), pp. 12562–12566. DOI: 10.1073/pnas.202427399.

[30] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. "Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method". In: *Physical Review Letters* 100.2 (Jan. 18, 2008), p. 020603. DOI: 10.1103/PhysRevLett.100.020603.

[31] Fangqiang Zhu and Gerhard Hummer. "Convergence and error estimation in free energy calculations using the weighted histogram analysis method". In: *Journal of Computational Chemistry* 33.4 (2012), pp. 453–465. DOI: 10.1002/jcc.21989.

[32] Shankar Kumar et al. "THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method". In: *Journal of Computational Chemistry* (1992). DOI: 10.1002/jcc.540130812.

[33] Maurice Karrenbrock et al. "Absolute Binding Free Energies with OneOPES". In: *The Journal of Physical Chemistry Letters* 15.39 (Oct. 3, 2024), pp. 9871–9880. DOI: 10.1021/acs.jpclett.4c02352.

[34] Anna S. Kamenik et al. "Peptidic Macrocycles - Conformational Sampling and Thermodynamic Characterization". In: *Journal of Chemical Information and Modeling* 58.5 (May 29, 2018), pp. 982–992. DOI: 10.1021/acs.jcim.8b00097.

[35] Gaurav Bhardwaj et al. "Accurate de novo design of membrane-traversing macrocycles". In: *Cell* 185.19 (Sept. 2022), 3520–3532.e26. DOI: 10.1016/j.cell.2022.07.019.

[36] Alessandro Laio and Francesco L. Gervasio. "Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science". In: *Reports on Progress in Physics* (2008). DOI: 10.1088/0034-4885/71/12/126601.

[37] Christopher D. Fu and Jim Pfaendtner. "Lifting the Curse of Dimensionality on Enhanced Sampling of Reaction Networks with Parallel Bias Metadynamics". In: *Journal of Chemical Theory and Computation* (2018). DOI: 10.1021/acs.jctc.7b01289.

[38] Omar Valsson, Pratyush Tiwary, and Michele Parrinello. "Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint". In: *Annual Review of Physical Chemistry* (2016). DOI: 10.1146/annurev-physchem-040215-112229.

[39] Pedro E.M. Lopes, Olgun Guvench, and Alexander D. MacKerell. "Current Status of Protein Force Fields for Molecular Dynamics". In: *Methods in molecular biology (Clifton, N.J.)* 1215 (2015), pp. 47–71. DOI: 10.1007/978-1-4939-1465-4_3.

[40] Paul S. Nerenberg and Teresa Head-Gordon. "Optimizing Protein-Solvent Force Fields to Reproduce Intrinsic Conformational Preferences of Model Peptides". In: *Journal of Chemical Theory and Computation* (2011). DOI: 10.1021/ct2000183.

[41] Stefano Piana, John L Klepeis, and David E Shaw. "Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations". In: *Current Opinion in Structural Biology* 24 (Feb. 2014), pp. 98–105. DOI: 10.1016/j.sbi.2013.12.006.

[42] Stefano Piana et al. "Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States". In: *The Journal of Physical Chemistry B* 119.16 (Apr. 23, 2015), pp. 5113–5123. DOI: 10.1021/jp508971m.

[43] Paul Robustelli, Stefano Piana, and David E. Shaw. "Developing a molecular dynamics force field for both folded and disordered protein states". In: *Proceedings of the National Academy of Sciences* 115.21 (May 22, 2018), E4758–E4766. DOI: 10.1073/pnas.1800690115.

[44] Stéphanie Jephthah et al. "Force Field Effects in Simulations of Flexible Peptides with Varying Polyproline II Propensity". In: *Journal of Chemical Theory and Computation* 17.10 (Oct. 12, 2021), pp. 6634–6646. DOI: 10.1021/acs.jctc.1c00408.

[45] Ferruccio Palazzesi et al. "Accuracy of Current All-Atom Force-Fields in Modeling Protein Disordered States". In: *Journal of Chemical Theory and Computation* (2014). DOI: 10.1021/ct500718s.

[46] Bhumika Singh, Yisel Martínez-Noa, and Alberto Perez. "How Well Do Molecular Dynamics Force Fields Model Peptides: A Systematic Benchmark across Diverse Folding Behaviors". In: *The Journal of Physical Chemistry B* 130.16 (Apr. 23, 2026), pp. 4344–4357. DOI: 10.1021/acs.jpcb.6c01176.

[47] Vittorio Limongelli, Massimiliano Bonomi, and Michele Parrinello. "Funnel metadynamics as accurate binding free-energy method". In: *Proceedings of the National Academy of Sciences* 110.16 (Apr. 16, 2013), pp. 6358–6363. DOI: 10.1073/pnas.1303186110.

[48] Valerio Rizzi et al. "OneOPES, a Combined Enhanced Sampling Method to Rule Them All". In: *Journal of Chemical Theory and Computation* 19.17 (Sept. 12, 2023), pp. 5731–5742. DOI: 10.1021/acs.jctc.3c00254.

[49] Giovanni Bussi. "Hamiltonian replica-exchange in GROMACS: a flexible implementation". In: *Molecular Physics* 112.3 (Feb. 16, 2014), pp. 379–384. DOI: 10.1080/00268976.2013.824126.

[50] Mark James Abraham et al. "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers". In: *SoftwareX* 1-2 (Sept. 2015), pp. 19–25. DOI: 10.1016/j.softx.2015.06.001.

[51] Gareth A. Tribello et al. "PLUMED 2: New feathers for an old bird". In: *Computer Physics Communications* 185.2 (Feb. 2014), pp. 604–613. DOI: 10.1016/j.cpc.2013.09.018.

[52] S. Pérez-Conesa, Pablo M. Piaggi, and Michele Parrinello. "A local fingerprint for hydrophobicity and hydrophilicity: From methane to peptides". In: *The Journal of Chemical Physics* (2019). DOI: 10.1063/1.5088418.

[53] Marc Schulze et al. "All you need is water: Converging ligand binding simulations with hydration collective variables". In: *The Journal of Chemical Physics* 163.16 (Oct. 23, 2025), p. 164111. DOI: 10.1063/5.0287856.

[54] Michele Invernizzi and Michele Parrinello. "Rethinking Metadynamics: From Bias Potentials to Probability Distributions". In: *The Journal of Physical Chemistry Letters* 11.7 (Apr. 2, 2020), pp. 2731–2736. DOI: 10.1021/acs.jpclett.0c00497.

[55] Michele Invernizzi, Pablo M. Piaggi, and Michele Parrinello. "Unified Approach to Enhanced Sampling". In: *Physical Review X* 10.4 (Nov. 17, 2020), p. 041034. DOI: 10.1103/PhysRevX.10.041034.

[56] Michele Invernizzi and Michele Parrinello. "Exploration vs Convergence Speed in Adaptive-Bias Enhanced Sampling". In: *Journal of Chemical Theory and Computation* 18.6 (June 14, 2022), pp. 3988–3996. DOI: 10.1021/acs.jctc.2c00152.

[57] Scott A. Hollingsworth and Ron O. Dror. "Molecular Dynamics Simulation for All". In: *Neuron* 99.6 (Sept. 2018), pp. 1129–1143. DOI: 10.1016/j.neuron.2018.08.011.

[58] Jin Zhang et al. "Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents". In: *Journal of Chemical Theory and Computation* 13.3 (Mar. 14, 2017), pp. 1034–1043. DOI: 10.1021/acs.jctc.7b00169.

[59] James A. Maier et al. "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB". In: *Journal of Chemical Theory and Computation* (2015). DOI: 10.1021/acs.jctc.5b00255.

[60] Luciano A. Abriata and Matteo Dal Peraro. "Assessment of transferable forcefields for protein simulations attests improved description of disordered states and secondary structure propensities, and hints at multi-protein systems as the next challenge for optimization". In: *Computational and Structural Biotechnology Journal* 19 (Jan. 2021), pp. 2626–2636. DOI: 10.1016/j.csbj.2021.04.050.

[61] Robert H. Swendsen and Jian-Sheng Wang. "Replica Monte Carlo Simulation of Spin-Glasses". In: *Physical Review Letters* 57.21 (Nov. 24, 1986), pp. 2607–2609. DOI: 10.1103/PhysRevLett.57.2607.

[62] Yuji Sugita and Yuko Okamoto. "Replica-exchange molecular dynamics method for protein folding". In: *Chemical Physics Letters* 314.1 (Nov. 1999), pp. 141–151. DOI: 10.1016/S0009-2614(99)01123-9.

[63] Ulrich H.E. Hansmann. "Parallel tempering algorithm for conformational studies of biological molecules". In: *Chemical Physics Letters* 281.1 (Dec. 1997), pp. 140–150. DOI: 10.1016/S0009-2614(97)01198-6.

[64] Lingle Wang, Richard A. Friesner, and B. J. Berne. "Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2)". In: *The Journal of Physical Chemistry B* (2011). DOI: 10.1021/jp204407d.

[65] Paula C. Ortet et al. "Recapitulating the Binding Affinity of Nrf2 for KEAP1 in a Cyclic Heptapeptide, Guided by NMR, X-ray Crystallography, and Machine Learning". In: *Journal of the American Chemical Society* 143.10 (Mar. 17, 2021), pp. 3779–3793. DOI: 10.1021/jacs.0c09799.

[66] S Luckett et al. "High-resolution structure of a potent, cyclic proteinase inhibitor from sunflower seeds". In: *Journal of Molecular Biology* 290.2 (July 1999), pp. 525–533. DOI: 10.1006/jmbi.1999.2891.

[67] Mengqi Zhong et al. “Interaction Energetics and Druggability of the Protein–Protein Interaction between Kelch-like ECH-Associated Protein 1 (KEAP1) and Nuclear Factor Erythroid 2 Like 2 (Nrf2)”. In: *Biochemistry* (2019). DOI: 10.1021/acs.biochem.9b00943.

[68] Luca Gagliardi and Walter Rocchia. “SiteFerret: Beyond Simple Pocket Identification in Proteins”. In: *Journal of Chemical Theory and Computation* 19.15 (Aug. 8, 2023), pp. 5242–5259. DOI: 10.1021/acs.jctc.2c01306.

[69] Giovanni Bocchi et al. “GENEOnet: a breakthrough in protein binding pocket detection using group equivariant non-expansive operators”. In: *Scientific Reports* 15.1 (Oct. 3, 2025), p. 34597. DOI: 10.1038/s41598-025-18132-5.

[70] Prashanth Choppara and Lokesh Bommareddy. “AI-driven protein pocket detection through integrating deep Q-networks for structural analysis”. In: *Journal of Computer-Aided Molecular Design* 39.1 (Oct. 6, 2025), p. 90. DOI: 10.1007/s10822-025-00669-2.

[71] Valerio Rizzi et al. “The role of water in host-guest interaction”. In: *Nature Communications* 12.1 (Jan. 4, 2021), p. 93. DOI: 10.1038/s41467-020-20310-0.

[72] Tamas Lazar et al. “Distance-Based Metrics for Comparing Conformational Ensembles of Intrinsically Disordered Proteins”. In: *Biophysical Journal* 118.12 (June 2020), pp. 2952–2965. DOI: 10.1016/j.bpj.2020.05.015.

[73] Gustavo Fuertes et al. “Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements”. In: *Proceedings of the National Academy of Sciences* (2017). DOI: 10.1073/pnas.1704692114.

[74] Emanuel Parzen. “On Estimation of a Probability Density Function and Mode”. In: *The Annals of Mathematical Statistics* (1962). DOI: 10.1214/aoms/1177704472.

[75] Martin Karplus and Joseph N. Kushick. “Method for estimating the configurational entropy of macromolecules”. In: *Macromolecules* (1981). DOI: 10.1021/ma50003a019.

[76] Yen-Chi Chen. *A Tutorial on Kernel Density Estimation and Recent Advances*. Sept. 12, 2017. DOI: 10.48550/arXiv.1704.03924.